Title: Unravelling genome-wide mosaic microsatellite mutations at single-cell resolution

Authors:

Chunyi Wang[1,2,9], Wenxuan Fan[1,2,9], Weixiang Wang[1,2,4,9], Yonghe Xia[1,2,9], Jinhong Lu[1,2,4,9], Xiaoyu Ma[1,2,9], Jichuan Yu[1,2,9], Yunchao Zheng[1,2], Yan Luo[1,2], Wenlong Li[1,2], Qing Yang[1,2,3], Meizhen Lin[1,2], Huan Liu[1,2], Yangning Lan[1,2], Chengyu Li[5], Xiaodong Liu[1,2], Danyang He[1,2], Shang Cai[1,2], Xiuyan Yu[6], Dan Zhou[7], Manolis Kellis[8], Xushen Xiong[5], Qi Xie[1,2*], and Yanmei Dou[1,2*]

1. School of Life Sciences, Westlake University, Hangzhou, Zhejiang, 310024, China
2. Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang, 310024, China
3. College of Life Sciences, Zhejiang University, Hangzhou, Zhejiang, China
4. Fudan University, Shanghai, 200433, China
5. Liangzhu Laboratory, Zhejiang University, Hangzhou, Zhejiang, China
6. Department of Breast Surgery, Second Affiliated Hospital, Zhejiang University, Hangzhou, Zhejiang, China
7. The Second Affiliated Hospital, School of Public Health, Zhejiang University School of Medicine, Hangzhou, China
8. Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
9. These authors contributed equally

*Correspondence: xieqi@westlake.edu.cn (Q.X.), douyanmei@westlake.edu.cn (Y.D.).

Abstract:

Short tandem repeats (STRs), or microsatellites, are highly mutable genomic elements that modulate gene regulations and are implicated in a range of human diseases. However, detecting mosaic STR mutations at single-cell resolution remains challenging due to both technical and biological complexities. To address this, we developed BayesMonSTR, a robust algorithm that enables accurate detection of mosaic STR mutations. Using this tool in single-cell analysis of human tissues, we reveal an accumulation of longer mosaic STR insertions and deletions (indels) in aging mitotic and post-mitotic cells. Strikingly, prefrontal cortex (PFC) neurons accumulate a higher burden of STR mutations than B cells or lung epithelium, with aged neurons exhibiting a particularly pronounced increase in longer STR deletions. These mutations are enriched at transcription start sites (TSSs) and active enhancers of highly expressed genes. Our work establishes a foundation for genome-wide, hypothesis-free discovery of disease-associated mosaic STR mutations and reveals a previously unexplored landscape of mosaic STR variation in development and aging.

Introduction:

Somatic mosaicism—postzygotic mutations acquired over a lifetime—reflects the cumulative interplay between genome instability, environmental exposures, and cellular repair processes. While mosaicism contributes to aging and diseases[1], its detection has focused predominantly on SNVs[2,3] and structural variants[4–6], overlooking STRs. STRs, spanning ~1-5% of the human genome, exhibit mutation rates up to 1,000-fold higher than non-repetitive regions due to polymerase slippage and replication errors[7]. Importantly, STRs regulate transcription, facilitate chromatin interactions, and are linked to nearly a hundred human disorders—a large proportion of which are neurological[8,9]. However, technical challenges—such as PCR artifacts, alignment ambiguities[7], and the inability to resolve STRs at single-cell resolution[10]—have obscured the landscape of mosaic STR mutations in aging tissues. Existing STR genotypers[11–13] are designed for germline variants or bulk tissues[14], failing to distinguish mosaic mutations from artifacts in single-cell data[15]. This gap has limited our ability to explore hypotheses about age-related STR genome instability and their roles in aging and disease.

Recognizing the critical gap in current methodologies, we developed BayesMonSTR, a computational framework that integrates probabilistic read segmentation, haplotype phasing, and a hierarchical Bayesian model to detect mosaic STR mutations with nucleotide-level resolution. By integrating comprehensive layers of data and quantifies all aspects in terms of probability, BayesMonSTR resolves ambiguities in extracting imperfect STR alleles (e.g., interruptions, truncated motifs) from noisy sequencing data, while rigorously filtering out technical artifacts and germline polymorphisms. Using this tool, we revealed the profiles of mosaic STR mutations in both post-mitotic and mitotic tissues during aging. Notably, neurons from healthy individuals accumulate age-associated long STR deletions, which are specifically enriched in the regulatory regions of highly expressed genes. Analysis of snATAC-seq data from the PFC neurons of Alzheimer's disease (AD) patients reveals a convergence of mosaic STR mutations in active regulatory regions of AD risk genes.

Our approach maps unexplored landscapes of mosaic STR mutations across aging tissues, enabling the genome-wide, hypothesis-free discovery of novel disease-associated variants. BayesMonSTR is open source, licensed under the MIT License, and publicly accessible on GitHub at https://github.com/douymLab/BayesMonSTR.

Results:

Overview of BayesMonSTR

BayesMonSTR is a computational framework designed for the robust detection of mosaic STR mutations from single cell sequencing data. The input to BayesMonSTR consists of sequencing reads and an STR panel, which defines a set of STR loci on the reference genome. For this study, we utilized a commonly adopted STR panel containing ~1.6 million loci[11], ensuring a relatively comprehensive coverage of the

genome's repetitive regions. Notably, unlike most existing methods that primarily estimate STR mutations length changes[16,17], BayesMonSTR provides nucleotide-level resolution, enabling not only the inference of STR length changes but also the identification of mosaic point mutations within the STR and its flanking regions. This capability significantly enhances the resolution and biological interpretability of STR mutation detection.

To process the sequencing reads, we employed a probabilistic hidden Markov model (HMM) for read segmentation, separating each read into STR and flanking regions (Fig. 1a, Methods). This read segmentation step ensures precise delineation of repetitive sequences from adjacent regions, reducing potential biases in downstream analyses, and the segmented reads were then used to generate sequence-based STR alleles (Supplementary Fig. 1).

The segmented reads were then used to generate haplotype-resolved STR alleles, a crucial step for distinguishing true STR mutations from sequencing artifacts. To enhance the accuracy of allele assignment, we implemented a probabilistic phasing framework (Fig. 1b, Methods). This framework integrates population haplotype information with sequencing reads from individual samples, enabling the calculation of linkage probabilities between STR alleles and nearby germline heterozygous single nucleotide polymorphisms (hSNPs).

BayesMonSTR then employs a hierarchical Bayesian network to jointly infer germline and somatic mosaic genotype posteriors integrating multiple layers of data, including population allele frequencies, bulk sequencing data and single cell sequencing data (Fig. 1c, Methods). After calculating the somatic posteriors for individual single cells, we applied a dynamic programming approach to estimate the probability that a given site contains at least one somatic cell within an individual (Methods). The key parameters of the Bayesian network were estimated using specialized methods based on real-world data. For example, stutter-error parameters were derived from a length-based stutter error model (Supplementary Fig. 2, Methods), while imbalanced allele frequencies in whole-genome amplified single-cell sequencing data were modeled using a piece-wised Gaussian process regression approach (Methods).

Following Bayesian genotype inference, over 50 carefully chosen features were extracted to distinguish mosaic mutations from artifacts (Fig. 1d, Methods). These features include metrics capturing read and base qualities, allele imbalances, and STR-specific error patterns, *etc.* (Supplementary Table 1). By generating a meticulously curated training set of over 10,000 orthogonally evaluated and manually checked candidate STR mutations from three different single-cell sequencing platforms (Supplementary Table 2), we employed a Random Forest (RF) classification model to identify mosaic STR mutations across the genome (Fig. 1e, Methods). Notably, we trained two separate models: one for phasable sites with at least one nearby hSNP and another for nonphasable sites, which we refer to as 'Phasable mode' and 'Nonphasable

mode'. In the RF model for Phasable mode, haplotype-phasing-related features were incorporated (Methods).

The comprehensive framework enables BayesMonSTR to achieve accurate detection of base pair changes within STR regions, including both indels and base substitutions. Figures 1f and 1g illustrate examples of mosaic STRs detected from two public datasets derived from two different sequencing platforms[18,19]. These visualizations, generated with our newly developed software BayesMonSTR-INSIGHT[20] specifically designed to provide advanced visualization of mosaic mutations, highlight BayesMonSTR's capability to accurately identify both indels (Fig. 1f) and mismatches (Fig. 1g).

Accurate detection of mosaic STR indels and mismatches from single cell sequencing data

Single cells possess only a single nucleus, which is used for sequencing, and mosaic mutations are typically present in a small fraction of cells or are cell specific. This makes it extremely challenging to assess the validity of mosaic mutations detected from real-world single cell sequencing data. To address this challenge, we constructed a 'gold-standard' benchmarking dataset, providing a more controlled and reliable approach for validation (Fig. 2a).

For this, we sequenced 28 single cells from a breast tissue sample, comprising six epithelial cells from the breast tumor tissue (Lin-EpCAM+CD49f+), 12 epithelial cells from the normal breast tissue (Lin-EpCAM+CD49f+), and ten Lin+ immune cells (Supplementary Fig. 3, Methods). Among these, nine cells were processed as duplicate libraries and sequenced twice, while 14 cells were subjected to both DNA and RNA sequencing from the same single cells. Single-cell DNA was amplified using two widely adopted strategies: MDA[21] and PTA[22] (Methods). Additionally, adjacent tissues were sequenced using bulk whole-genome sequencing (WGS). This comprehensive approach allowed us to generate three distinct types of orthogonal benchmarking data: (1) duplicate cell libraries, (2) single-cell DNA and RNA sequencing, and (3) bulk WGS sequencing from adjacent tissues. Figure 2B illustrates a validated site using the three orthogonal approaches.

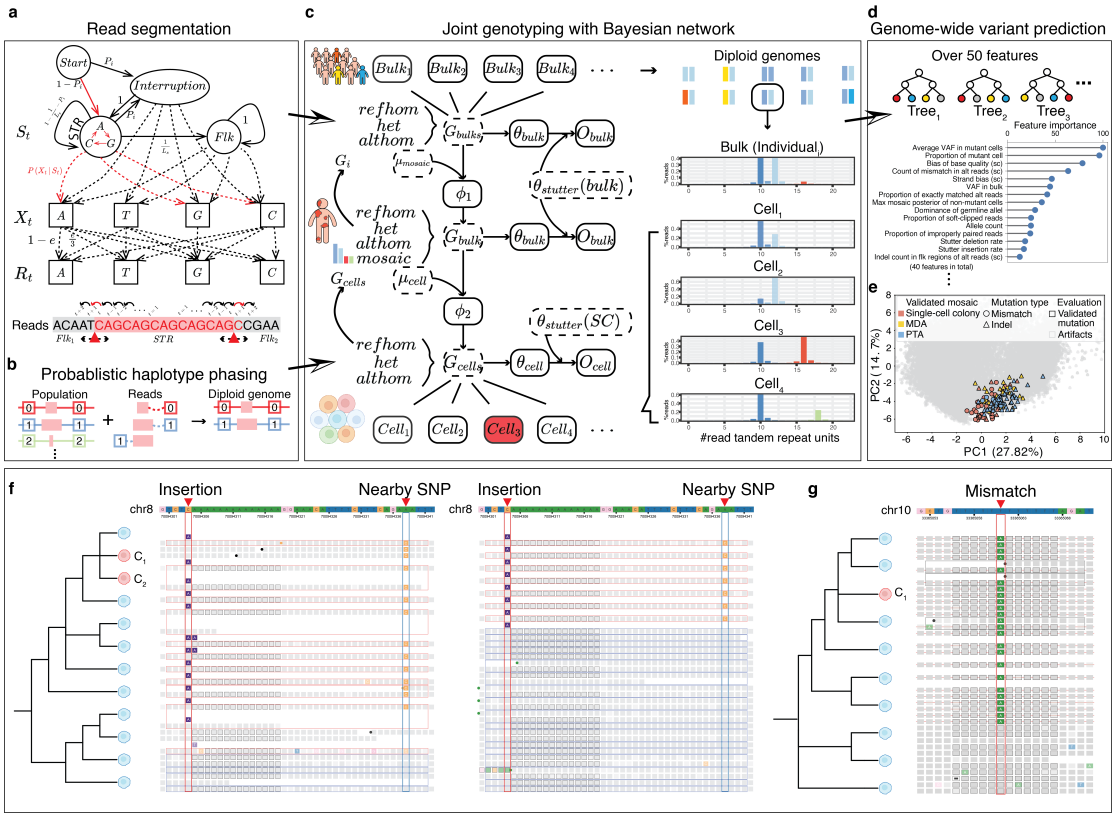


Fig. 1 Overview of BayesMonSTR. (**a**) BayesMonSTR employs an HMM to probabilistically segment each sequencing read, distinguishing the STR region from its flanking sequences. (**b**) STR alleles are assigned to their chromosomal haplotypes by statistically linking them to the alleles of neighboring heterozygous variants. (**c**) A Hierarchical Bayesian Network integrates three layers of data—population-level sequencing, individual-level bulk sequencing, and single cell sequencing—to calculate individual- and cell-level genotype posteriors. (**d**) An RF model, trained on over 50 carefully chosen read-level and genome-level features, classifies true mosaic mutations from artifacts. (**e**) PCA constructed from top 40 features shows that validated mosaic variants from independent single-cell sequencing platforms cluster in similar regions, highlighting the robustness of BayesMonSTR. (**f**) Illustration of a mosaic 1-bp insertion. A 1-bp insertion mutation detected in single-cell colony data is shown. Supporting sequencing reads covering the variant site are displayed for two distinct cells harboring the mutation. (**g**) A cell-specific base substitution within a short tandem repeat (STR) region, identified from single-cell whole-genome sequencing (scWGS) data.

The performance of BayesMonSTR was evaluated against HipSTR[11] (version 0.6.2) and ExpanshionHunter[12] (version 5.0.0). Notably, we adapted the state-of-the-art germline STR callers for single cell sequencing data. By applying a series of hard filters (Methods), we enabled these tools to detect mosaic STR mutations.

Figures 2c–g presents the results generated using the gold-standard benchmarking dataset described above. Across these tests, we demonstrate that BayesMonSTR

achieves a remarkable multi-fold improvement in precision for detecting mosaic STR mutations (Fig. 2c, Supplementary Fig. 4a-b, Supplementary Table 3-4). Specifically, BayesMonSTR accurately identifies both cell-specific and shared STR mutations in both tumor-tissue cells and normal-tissue cells, encompassing both indels and base substitutions. One notable observation is that the accuracy of mosaic detection from PTA data was generally lower, likely due to the 9-10 rounds of PCR involved in the PTA library preparation stage (Methods), which introduced additional DNA-slippage errors (Supplementary Fig. 2).

To better understand the superior performance of BayesMonSTR compared to other software tools, we thoroughly examined artifact variants generated by different tools. Our analysis (Fig. 2d-f, Supplementary Fig. 4c-g) revealed that BayesMonSTR achieved a significant improvement in accuracy due to its comprehensive framework, which includes the joint genotyping with stutter-error models, effective probabilistic haplotype phasing, and a series of informative features integrated into the RF classification model.

As mentioned above, the hypermutable nature of STR regions makes them promising targets for *in vivo* lineage tracing. By leveraging predicted mosaic STR mutations shared across multiple cells, we constructed a lineage tree[23] (Supplementary Table 5, Methods). The tree generated from STR mutations across ~1% of the genome resembles the topology of the tree constructed from SNVs detected in the remaining 99% of the genome (Fig. 2g, Methods), supporting the potential using STR mutations for cost-efficient lineage tracing.

To further evaluate the robustness of BayesMonSTR, we analyzed three publicly available single-cell sequencing datasets generated using diverse strategies, including single-cell whole genome amplification (WGA) and single-cell colonies[18,24–26]. These datasets were specifically chosen based on the availability of orthogonal data for cross validation, such as duplicate single-cell libraries or adjacent tissue samples. Across multiple benchmarks, BayesMonSTR consistently achieved a superior accuracy for detecting mosaic STR mutations, underscoring its robustness (Supplementary Fig. 4h). For example, in the single-cell colony dataset comprising cells from adjacent skin samples, BayesMonSTR identified the highest number of shared variants between the cells from adjacent skin samples, compared to remote tissues, further validating its performance (Fig. 2h).

Finally, we evaluated BayesMonSTR on simulated data with spike-in mutations (Supplementary Fig. 5, Supplementary Table 6, Methods). This analysis demonstrated that BayesMonSTR achieved substantially higher precision than other tools by effectively excluding most artifacts and germline variants, while maintaining a competitive sensitivity profile (0.4-0.6 vs. 0.6-0.8).

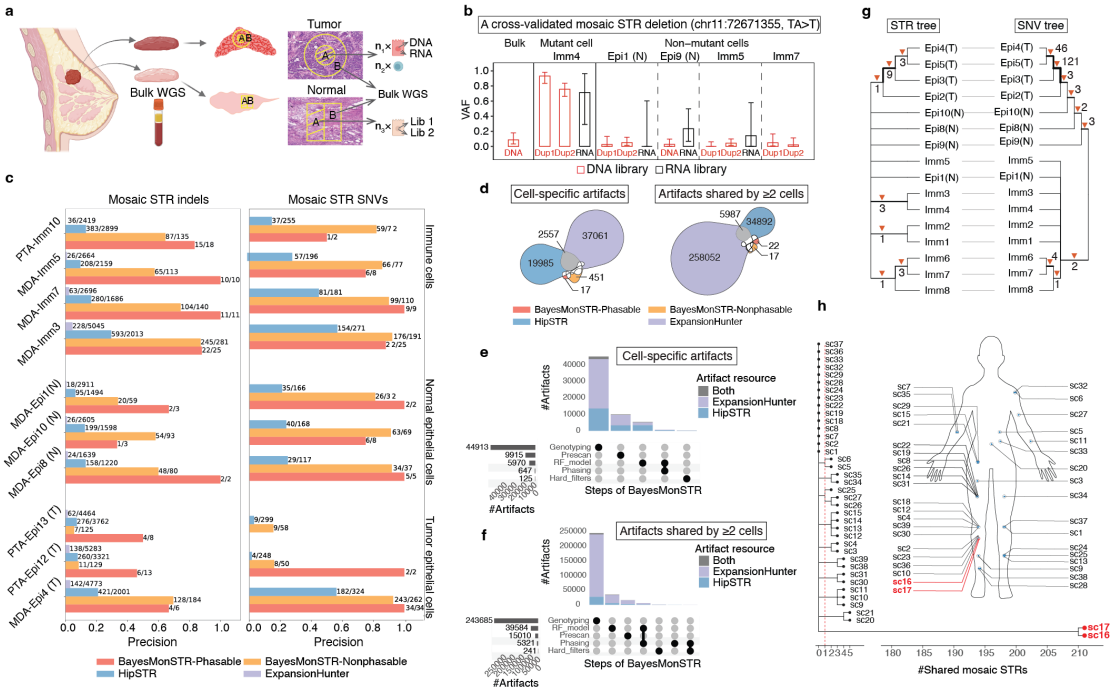


Fig. 2 Accurate detection of mosaic STR indels and SNVs from single cell sequencing data. (**a**) To construct a gold-standard benchmarking dataset, we used 28 single cells from a breast tissue sample, comprising six epithelial cells (Epi) from tumor tissue (T), 12 epithelial cells from normal tissue (N), and 10 immune (Imm) cells. Of these, 12 cells had duplicate libraries and were sequenced twice, while 15 cells underwent both DNA and RNA sequencing from the same single cell. Single cell DNA was amplified using two widely adopted strategies, PTA and MDA. (**b**) An illustrative site validated through both duplicate library sequencing, and RNA-seq derived from the same single cell. The mutant reads were also present in the bulk DNA-seq from the adjacent tissue. Error bars represent 95% confidence intervals of the mutant-cell VAFs calculated using binomial sampling. (**c**) BayesMonSTR achieved a multi-fold improvement in detecting cell-specific mosaic STR mutations, compared to the adapted versions of HipSTR and ExpansionHunter. (**d-f**) For both cell-specific and shared variants, BayesMonSTR demonstrated a significant improvement in accuracy, largely due to its comprehensive framework, which includes effective probabilistic read segmentation, haplotype phasing, joint genotyping, and the incorporation of informative features into the RF classification model. In panels **e-f,** solid points represent artifact mutations called by other tools that were filtered out by BayesMonSTR's genotyping, RF prediction model or hard filters. (**g**) The phylogenetic tree reconstructed using mosaic STR mutations resembles the tree built from point mutations, while providing enhanced resolution in certain branches. The width of the branch was positively correlated with the number of mutations that arise at the branching point. Numbers indicate the specific number of mutations that arose at each branching point. (**h**) From a public single cell colony dataset which includes two clones from adjacent skin samples as well as samples with multiple clones from remote tissues, BayesMonSTR detected the highest number of shared STR mutations in the two clones

237 from the adjacent skin samples. The human figure was adapted from the image
238 available at https://scv.cancervision.com/Screen1/DB10.

239

240 **Landscape of mosaic STRs from mitotic and post-mitotic cells**

241 Genome instability arises from a complex interplay of biological and chemical
242 processes, including replication stress, DNA repair activity, transcription-associated
243 DNA damage and repair, endogenous metabolic byproducts, exogenous environmental
244 factors, *etc*. To investigate STR-associated instability in different contexts, we analyzed
245 publicly available datasets from three distinct cell types with varying proliferative
246 potentials and environmental exposures: (1) B cells[27] from the blood, which undergo
247 frequent replication during their lifespan ; (2) lung epithelial cells[28], which are
248 proliferative particularly during the wound-healing process, and (3) neurons[26], which
249 cease dividing at an early developmental stage and are highly transcriptionally active.

250 From the collected datasets, BayesMonSTR identified 4,008 mosaic STR indels and
251 971 mosaic STR SNVs (defined as SNVs within or flanking STRs by ±5 bp) from 56
252 B cells across 14 individuals. Additionally, it identified 1,745 mosaic STR indels and
253 3,447 STR SNVs from 160 lung epithelial single-cell colonies across 16 individuals,
254 and 13,488 mosaic STR indels and 2,353 STR SNVs from 126 neurons across 19
255 individuals (Supplementary Table 7). The mutation rate per cell per base pair ranges
256 from approximately 10^{-6} to 10^{-5}, significantly higher than the reported point mutation
257 rate outside of microsatellite regions[2]. The mutation signatures differ across tissues
258 (Methods, Supplementary Fig. 6), and the overall exonic indel mutation rate was about
259 60% lower than expected ($p = 1.08\times10^{-121}$, two-tailed binomial test, Table 1,
260 Supplementary Table 8), suggesting a strong purifying selection.

261

Table 1 Likely pathogenic exonic STR mutations identified in this study

Pos	Mutation	Cell type	Individual	Gene	Mutation type	CADD score	Polyphen2 score	Associated Disease
chr3:195254670	CT>C	B cell	A11365	*PPP1R2*	frameshift	6.348		Alzheimer’s disease
chr5:45695972	TCGC>T	B cell	A19553	*HCN1*	inframe	17.83		Epilepsy
chr6:109906329	GCTT>G	B cell	BF2	*AK9*	inframe	15.01		Developmental delay
chr12:6777069	T>TTGC	B cell	BF2	*ZNF384*	inframe	6.415		Acute myeloid leukemia
chr1:87045896	TCCTACA>T	B cell	BM2	*CLCA4*	inframe	1.747		Cystic fibrosis
chr3:32022616	C>T	B cell	A19552	*OSBPL10*	nonsynonymous	25.1	0.776	Liver disease
chr12:54332796	G>A	B cell	A19553	*HOXC13*	nonsynonymous	16.01	0.949	Ectodermal dysplasia
chr19:48342606	CA>C	Neuron	UMB5559	*CRX*	frameshift	24.4		Retinitis pigmentosa
chr3:19554535	A>AG	Neuron	UMB5532	*KCNH8*	frameshift	27.5		Epilepsy
chr19:30503438	G>GGT	Neuron	UMB5823	*URI1*	frameshift	30		Intellectual disability
chr4:47945399	TTTTTTC>T	Neuron	UMB5871	*CNGA1*	inframe	18.74		Retinitis pigmentosa
chr3:65425560	T>TCTG	Neuron	UMB5823	*MAGI1*	inframe	14.23		Autism spectrum disorder
chr2:168104768	T>A	Neuron	UMB5511	*XIRP2*	nonsynonymous	22.9	0.986	Muscular dystrophy
chr18:21594903	C>G	Neuron	UMB5657	*TTC39C*	nonsynonymous	21.2	0.821	Developmental delay
chr14:65392796	C>A	Neuron	UMB1278	*CHURC1*	nonsynonymous	32		Autism spectrum disorder
chr19:35940786	T>A	Neuron	UMB1465	*FFAR2*	nonsynonymous	27.9	1	Type 2 diabetes
chr6:127837528	GCT>G	Lung epithelium	PD34207	*SOGA3*	frameshift	24.3		Autism spectrum disorder
chr5:14487617	$(GGC)_8A>G$	Lung epithelium	PD34207	*TRIO*	inframe	18.91		Autism spectrum disorder
chr12:51693392	T>G	Lung epithelium	PD34209	*BIN2*	nonsynonymous	16.2085	0.977	Cardiomyopathy

262 **Age-related accumulation of longer STR indels**

263 We observed that the total burden of mosaic mutations within and flanking STR
264 regions—including both indels and point mutations—increases with age in B cells
265 (p=0.002, one-tailed t test) and lung epithelial cells ($p=1.7 \times 10^{-10}$, one-tailed t test). In
266 contrast, mutation burden remains relatively stable in post-mitotic neurons across
267 donors of different ages (p=0.198, one-tailed t test, Fig. 3a). This is consistent with the
268 established understanding that STR-associated mutations arise primarily from
269 replication slippage during DNA proliferation. Notably, neurons exhibited a
270 significantly higher mosaic mutation burden than the other two cell types (vs. B cells:
271 $p = 1.3 \times 10^{-34}$; vs. lung epithelium: $p = 1.9 \times 10^{-5}$; two-tailed Mann–Whitney U test;
272 Fig. 3b), suggesting that mosaic STR mutations in neurons accumulate rapidly during
273 early development.

274 Indels of different lengths arise from distinct mutational mechanisms. Longer indels
275 affect more genomic sequence, therefore posing a higher risk of DNA structure
276 disruption. To determine whether cells in healthy individuals accumulate longer STR
277 indels with age, we analyzed indel length distributions across three cell types.
278 Interestingly, both indel length (Fig. 3c) and the burden of long deletions
279 (Supplementary Fig. 7a) increased significantly with age in lung epithelium, B cells,
280 and neurons. Specific p-values for the increase in indel length were: lung epithelium
281 (p=0.004), B cells (p=0.019), and neurons ($p=9.6 \times 10^{-4}$; one-tailed t-test); p-values for
282 the increased long-deletion mutation rates were: lung epithelium (p=0.023), B cells
283 (p=0.091), and neurons ($p=8.7 \times 10^{-4}$; one-tailed t-test). Full mutation list is provided
284 in Supplementary Table 9. This trend was particularly striking in neurons, as their
285 burden of 1-bp STR mutations shows no age-related increase (p=0.456, one-tailed t test,
286 Supplementary Fig. 7b), suggesting that longer STR indels are generated via distinct,
287 age-dependent mechanisms, which shape a unique mutational spectrum over the
288 neuronal lifespan.

289 These longer STR indels are prevalent across neurons. By the age of 80, approximately
290 85% of neurons from healthy individuals carry at least one STR indel greater than 5 bp
291 (Fig. 3d, Methods). Across the 126 neurons from 19 individuals, while most cells harbor
292 less than ten such STR indels, some cells contain more (Fig. 3e).

293 Notably, one aged neuron from an 80-year-old female exhibited nearly 50 long STR
294 indels. Intriguingly, this neuron—which shows a high burden of long STR indels—
295 carries a nonsynonymous mosaic exonic SNV in *PALB2* (Partner and Localizer of
296 BRCA2, Supplementary Table 10, Methods), an essential gene involved in DNA
297 repair[29]. This cell harbored a >30-fold higher number of >5bp mosaic STR indels than
298 other two neurons from the same individual (Fig. 3f), suggesting a possible defect in
299 microsatellite repair driven by mosaic SNVs in DNA repair–related genes.

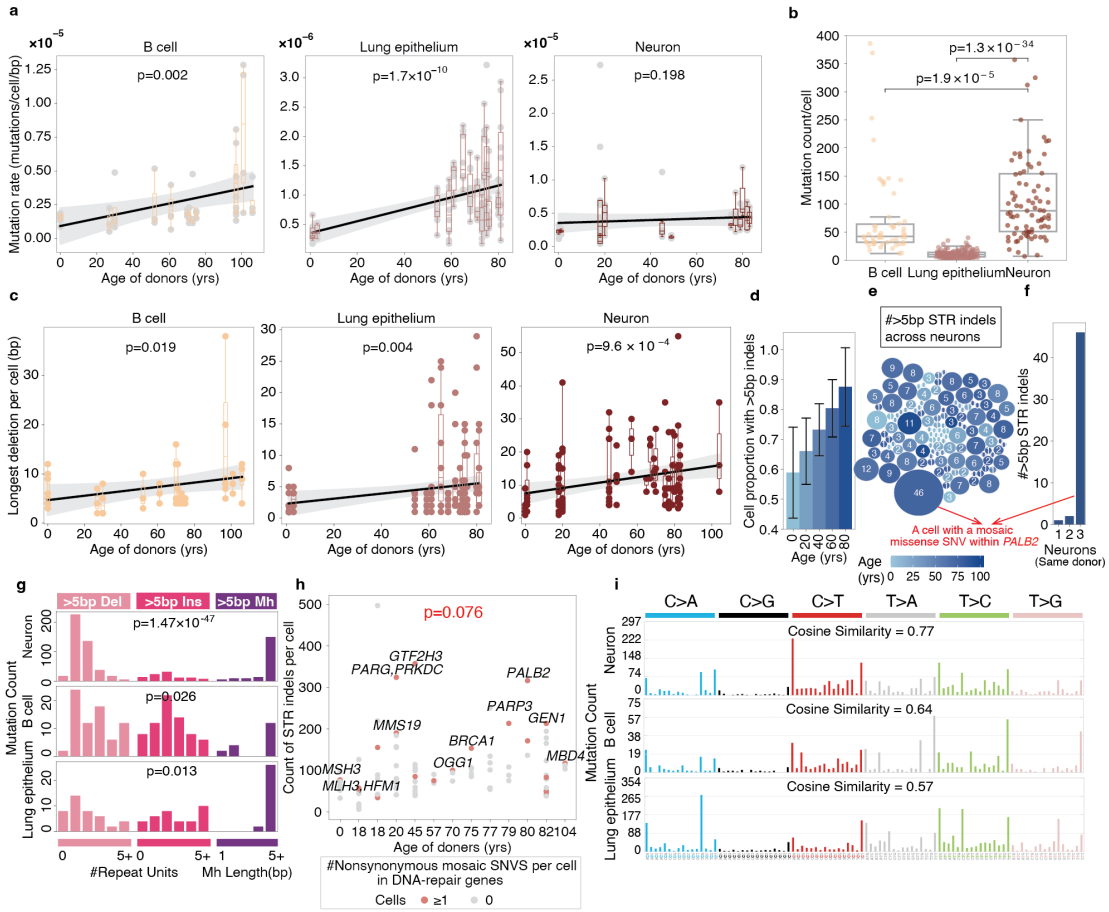


Fig. 3 Mosaic STR mutations by age across tissues and mutation mechanism analysis. (**a**) Mosaic mutations within and flank STR regions, including both indels and point mutations, accumulate with age in B cells and lung epithelial cells. In contrast, the mutation rate remains relatively stable in non-dividing neurons across donors of different ages. Linear regression equations were derived using the least squares method. P-values were calculated using a one-tailed t-test to test whether the regression coefficient is significantly greater than zero. The shaded areas around each curve represent the 95% CIs, assuming the differences between predicted values and observed values follow a t-distribution. Due to higher stutter errors in PTA-amplified cells, 1-bp STR indels from PTA cells showed relatively low validation rates, so only MDA-amplified neurons were used in panel a. Longer STR indels, which have higher signal-to-noise ratios, were included in the analysis of panels c-j. All long indels passed manual validation. (**b**) The total mosaic mutation burden per cell in neurons was significantly higher compared to the other two cell types. P-values were calculated using a two-tailed Mann-Whitney U test. (**c**) Across all three cell types, the length of longest mosaic STR contraction per cell increased significantly by age. The regressions, p-values, and CIs were calculated using the same approach as in panel A. (**d**) Estimated proportion of neurons carrying mosaic STR indels greater than 5bp by age. The proportion of neurons was predicted by a linear regression. Error bars indicate the 95% CIs assuming the differences between predicted values and observed values follow a t-

distribution. P-values were calculated with a two-tailed Mann-Whitney U test. All insertions greater than 5 bp included in panels d-g have passed the manual verification process. (**e**) Mosaic STR indels greater than 5 bp were prevalent across neurons, particularly in aged neurons. Each circles represents one individual neuron, with the numbers inside each circle indicating the count of >5bp STR indels acquired by each neuron. The gradient color scale represents the varying ages of the donors. The neuron carrying 46 >5 bp STR indels also harbors a missense mutation in the DNA repair gene *PLAB2*. (**f**) Distribution of >5 bp STR indels across three neurons from the same individual. The cell (same with the cell in panel e that carries 46 long STR indels) with the *PLAB2* missense mutation exhibits a substantially higher mutational burden of >5bp STR indels than the other two neurons. (**g**) Mutation signatures of >5bp STR indels across three cell types. Mh refers to microhomology. (**h**) The presence of nonsynonymous mosaic SNVs in DNA repair genes was marginally significantly associated with an increased count of mosaic STR indels in neurons. Each point represents one cell, and red points indicate cells carrying at least one nonsynonymous mosaic SNV in DNA-repair related genes. See Methods for details. Analyses were performed using GLMM with a negative binomial distribution, including a random intercept for individual to account for within-individual correlations. Critical genes associated with DNA repair were annotated beside the mutant cells. (**i**) Three-nucleotide context mutation signature profile of STR-region SNVs across three cell types. The numbers in the figure represent the cosine similarity between each mutation profile and the established patterns of genome-wide mosaic SNVs in prefrontal cortex (PMID 41166474).

Intrinsic microsatellite mutagenesis in aging tissues

To investigate the mechanisms underlying longer STR indels, we analyzed the properties of the microsatellites in which they occur, focusing on tract length, motif composition, and genomic location. This analysis revealed that longer indels preferentially occur in STRs that are themselves longer ($p=1.9\times10^{-21}$, two-tailed Mann-Whitney U test, Supplementary Fig. 8a) and have higher GC content ($p=1.0\times10^{-32}$, two-tailed Kolmogorov-Smirnov test, Supplementary Fig. 8b). We further conducted mutation signature analyses (Methods). The indel signatures of >5bp STR indels across all three cell types resembled those observed in *MSH2* (MutS homolog 2) and *MSH3* (MutS homolog 3) gene knockouts[30] (Supplementary Fig. 8c), implicating potential deficiencies in human MutSβ, which normally binds unpaired DNA loops commonly found in STR regions[31].

Notably, while >5 bp insertions and deletions occurred at similar frequencies in B cells and lung epithelium, neurons displayed a significant bias toward deletions over insertions ($p=1.47\times10^{-47}$, one-tailed Binomial test, Fig. 3f). Among these deletions, 29.8% were flanked by ≥5 bp microhomology—a hallmark of microhomology-mediated end joining (MMEJ), which typically generates long deletions. This pattern

361 suggests increased double-strand break formation and preferential MMEJ repair in
362 aging neurons[32].

363 To further investigate whether STR indel burden correlates with DNA repair capacity,
364 we examined the relationship between STR indel counts and the presence of
365 nonsynonymous exonic mosaic SNVs in DNA repair genes across neurons
366 (Supplementary Table 10). Using a negative binomial generalized linear mixed model
367 (GLMM) with individual as a random effect to account for donor-level variability, we
368 tested for a positive correlation between per-cell STR indel burden and the presence of
369 such mutations. Neurons carrying nonsynonymous SNVs in DNA repair genes showed
370 a marginally significant increase in STR indel burden ($p = 0.076$, negative binomial
371 GLMM, Fig. 3g, Methods). Intriguingly, among the neurons from each donor, several
372 exhibited substantially higher STR indel counts than the others from the same
373 individual. Notably, these outliers harbored nonsynonymous mutations in key DNA
374 repair genes, mainly including several double strand break repair genes (*PRKDC,*
375 *PARP3, BRCA1, PALB2, GEN1*)[33] and *GTF2H3*, a core factor in nucleotide excision
376 repair (NER)[34]. Together, these results further support the link between neuronal STR
377 indels and age-related DNA repair deficiencies driven by genomic instability.

378 Beyond indels, we further examined the signatures of mosaic SNVs within STR regions.
379 Notably, STR-region SNVs largely mirrored the established patterns of extra-STR
380 mosaic SNVs—attributed mainly to oxidative damage and cytosine deamination[35] (Fig.
381 3i, Supplementary Fig. 8d). Together with our findings on indel signatures and repair-
382 pathway involvement, these results paint a complex picture of microsatellite
383 mutagenesis in aging, marked by both age-associated genomic instability and the
384 engagement of specific DNA repair processes.

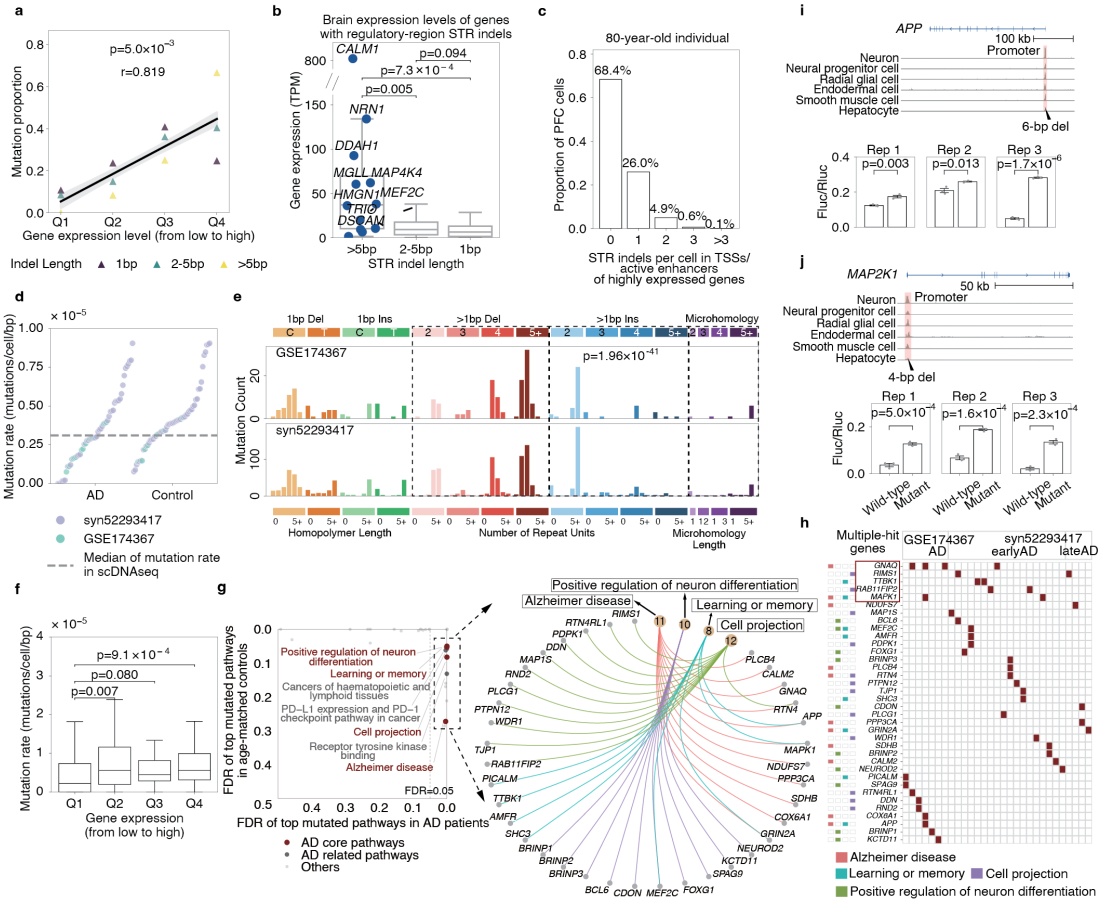


Fig. 4 Regulatory mosaic STR indels in neurons. (**a**) Mosaic STR indels in neurons are enriched in TSSs and active enhancers of highly expressed genes. The number of variants is indicated beside each data point. Q1-Q4 represent expression level quantiles from low to high (Methods), and the brain gene expression data were obtained from ENCODE. Linear regression equations were derived using the least squares method. P-values were calculated using a one-tailed t-test to test whether the regression coefficient is significantly greater than zero. The shaded areas around each curve represent the 95% CIs, assuming the differences between predicted values and observed values follow a t-distribution. (**b**) Genes with >5 bp mosaic STR indels in the TSS regions or active enhancers exhibit elevated gene expression levels, compared to genes with shorter indels in the TSS regions or active enhancers. P-value was calculated using a two-tailed Mann-Whitney U test. (**c**) Based on Poisson modeling—using the product of the total length of regulatory regions and the median mutation rate as the parameter μ—we estimate that over 30% of neurons in an 80-year-old individual carry at least one mosaic STR indel within the TSSs or active enhancers of highly expressed genes. (**d**) snATAC-seq samples exhibit mutation rates highly consistent with those observed in scWGS data. (**e**) A significant excess of >1 bp deletions over >1 bp insertions was observed in snATAC-seq data. P-value was calculated using a one-tailed Binomial test. (**f**) The STR mutation rate in active regulatory regions is increased in more highly expressed genes. P-values were calculated with a two-tailed t-test. (**g**) In AD patients, but not in age-

matched healthy controls, STR indels within transcription start sites (TSSs) and active enhancers were significantly enriched for genes in AD-related pathways (Methods). (**h**) Multiple key genes in AD pathways contained multi-hit STR indels in their active regulatory regions across different patients. "earlyAD" and "lateAD" indicate early-stage AD patients and late-stage AD patients. (**i-j**) Functional validation of somatic STR indels in regulatory regions using luciferase assays. Normalized levels of H3K4me3 and H3K27ac—key histone modifications linked to promoter and enhancer activities—were obtained from the ENCODE project. Pink regions indicate the promoter or enhancer regions of the gene, with indels highlighted by black triangles. Luciferase assays in HEK293T cells revealed that a 6-bp deletion in the TSS of *APP* and a 4-bp deletion in the TSS of *MAP2K1* significantly enhanced gene expression ($P < 0.05$, two-tailed t-test), with error bars representing standard deviation (3 biological replicates).

Mosaic STR indels in neuronal open chromatin

Given the established involvement of microsatellites in gene regulation[36], we investigated the distribution of STR indels in active regulatory regions across neurons. We found these mutations are more enriched in the TSS and active enhancers of genes with high expression levels in the brain ($p=5.0\times10^{-3}$, one-tailed t-test, Fig. 4a). Notably, longer STR indels showed the strongest enrichment (Fig. 4b). Given the mutation prevalence, over 30% of neurons in an 80-year-old individual would harbor at least one mosaic STR indels within TSSs of active enhancers of highly expressed genes based on Poisson sampling (Fig. 4c).

To further explore the enrichment of STR mutations in neuronal open chromatin regions, we next adapted the BayesMonSTR framework to detect mosaic STR indels directly from snATAC-seq data (Methods). Based on snATAC-seq data from the prefrontal cortex (PFC) of 55 AD patients and 55 age-matched healthy controls from two independent datasets[37,38], the mutation rate closely matched that observed in sc-WGS data (Fig. 4d, Methods). Like patterns in neuronal scWGS data, we observed a significant excess of long deletions over long insertions ($p=1.96\times10^{-41}$, one-tailed Binomial test, Fig. 4e). Crucially, the STR mutation rate was elevated in the active regulatory regions of highly expressed genes, mirroring the pattern we observed in scWGS data (Fig. 4f).

Intriguingly, while the per-cell mutation rate was similar between AD patients and controls (Fig. 4d), STR indels in AD patients located within TSSs and active enhancers were significantly enriched in genes associated with AD-related pathways (Fig. 4g-h, Methods). These included ontologies and genes related to learning and memory, amyloid-β (Aβ) pathway and synaptic function, positive regulation of neuron differentiation, cell projection, etc. (Supplementary Table 11). Many of the affected genes are well-recognized AD risk genes, such *APP* (amyloid beta precursor protein)[39], *CD2AP* (CD2-associated protein)[40], *PICALM* (phosphatidylinositol binding clathrin

446 assembly protein)[41], *MAPK1* (mitogen-activated protein kinase 1)[42], *MEF2C*
447 (Myocyte-specific enhancer factor 2C)[43], etc.

448 To test the function of these mutations, we further conducted luciferase assays of six
449 STR indels within the TSS regions and active enhancers of *APP, CD2AP, MAP2K1,*
450 *PLCB4, BRINP,* and *PPP3A* (Methods). Remarkably, experimental results showed that
451 4 out of 6 mutations led to significantly disrupted gene expressions in 3 replicated
452 experiments in 293T cells (Fig. 4i-j, Supplementary Fig. 9). These findings indicate that
453 microsatellites in active regulatory regions of highly brain-expressed genes are
454 particularly mutation-prone and may affect gene expression.

455 **Discussion**

456 Short tandem repeats are vastly overrepresented in the human genome, occurring far
457 more frequently than what would be expected by chance[44]. Serving as valuable genomic
458 elements for fast evolving, the hypermutable nature of STRs also makes them
459 vulnerable to harmful mutations. The single-cell genomic landscape of mosaic STR
460 mutations has not been explored before due to technical difficulties and the complexity
461 of these sequences. In this study, we present BayesMonSTR, the first tool to accurately
462 detect genome-wide somatic STR mutations in single cells. This capability provides a
463 unique opportunity to map the unexplored dynamic landscape of STR mutagenesis and
464 enables the hypothesis-free discovery of mosaic STR mutations associated with aging
465 and disease.

466 By characterizing mosaic STR mutations across mitotic and post-mitotic tissues, we
467 uncovered heterogeneous mutation rates and profiles. Unexpectedly, the mutation rate
468 in post-mitotic neurons was significantly higher than in mitotic B cells and lung
469 epithelial cells from age-matched individuals. The accumulation of longer STR indels
470 in neurons with aging was especially intriguing. Notably, we observed a neuron-
471 specific increase in longer STR deletions versus insertions over time, a pattern distinct
472 from that seen in mitotic B cells and lung epithelium. Mechanistic analysis further
473 suggested that these mutations may arise from double-strand DNA breaks and coupled
474 MMEJ. Furthermore, we found that longer STR indels accumulating in neurons with
475 age were enriched in the active regulatory regions of highly expressed genes; In
476 published snATAC-seq datasets from AD patients PFC cells, these mosaic STR indels
477 were significantly enriched in established AD risk genes and associated pathways.

478 Finally, BayesMonSTR provides a versatile, unified toolkit for detecting mosaic STR
479 mutations from single-cell WGS, snATAC-seq (BayesMonSTR-ATAC), and bulk
480 WGS data (BayesMonSTR-BulkMonSTR)[45], as well as enhanced visualization of STR
481 mutations (BayesMonSTR-INSIGHT)[20]. By enabling precise characterization across
482 these platforms, it opens new opportunities to investigate the contribution of STR
483 mutations to aging and disease.

484 **Acknowledgements** We are deeply grateful to the generous tissue contributions. We are
485 grateful to Dr. Peter J. Park for his critical insights and discussions; Drs. Xushen Xiong,

Manolis Kellis, Xiaodong Liu, Jing Zhang, Shumin Duan and Dan Zhou for data sharing and discussions; and Drs. Shang Cai, Danyang He, Hongtao Yu, Xiaochun Yu, Wei Wu, Lovelace J. Luquette, Ge Gao, Kai Zhang, Yanxiao Zhang, Jia Zheng, Yuxiang Huang, and Dangsheng Li for their comments and assistance. This work was supported by the Westlake Laboratory of Life Sciences and Biomedicine (Hangzhou 310024, Zhejiang, China), under the grant "Key R&D Program of Zhejiang" (2024SSYS0032) as well as the National Natural Science Foundation of China (32270682) to Y.D.; and the National Natural Science Foundation of China (82525055 and 82473121) to Q.X. We thank the Flow Cytometry Core Facility and the Genomics Core Facility at Westlake University for their experimental assistance, and the High-Performance Computing Center for technical support. We would like to thank the submitters of the following datasets: dbGaP accession phs001808.v1.p1, dbGaP accession phs001485.v3.p1, the European Genome-phenome Archive (EGA) accession EGAD00001005193, the EGA accession EGAS00001001934 and Synapse accession syn52293417. Data used in this study under the dbGaP accession phs001485.v3.p1 was generated by Drs. Christopher A. Walsh and Peter J. Park with funding from NINDS grant R01NS032457. The public datasets were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap and the EGA at https://ega-archive.org/.

Author contribution Y.D. conceived and supervised the project and secured the funding. Q.X., D.H., X.L., S.C., X.X., M.K. provided support in the supervision of the project. C.W., W.F., W.W., Y.X., J.L., X.M. and J.Y. all made significant contributions to the project. W.F. contributed to software development with help from C.W., W.W., Y.X., J.L. and Y.Z., C.W. developed the pipeline analyzing snATACseq data and lead the bioinformatics analysis. Y.X contributed to phylogeny reconstruction, public data downloading and developed the mutation-visulization software. Q.Y. helped phylogeny reconstruction. J.L. was responsible for generating the benchmarking dataset with help from X.Y., X.M., Y.L., and Z.M., X.Y. performed the breast tumor surgery. J.Y. and X.M. performed the functional assays. N.Y. contributed to cell sorting and related wet lab experiments. Y.Z. and J.L. were responsible for generating non-STR SNV mutation callings. W.L, W.W., C.W., Y.X, W.F. and H.L. contributed to mutation examination. Y.D., C.W., W.F., W.W. and J.L. wrote the manuscript. All authors carefully reviewed and approved the final manuscript.

Inclusion & Ethics statement This study was conducted in accordance with ethical guidelines and principles. It involved one human participant, and no animal subjects were included. The authors affirm that all research procedures were performed with respect for the participant's rights and well-being. The sequencing data generated in this study are publicly available, and no personally identifiable information was included. The authors ensure that the study adheres to the highest ethical standards in research, data generation, and usage.

Declaration of generative AI and AI-assisted technologies During the preparation of this work the authors used DeepSeek in order to polish our language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Competing interests The authors declare no competing interests.

527 **References**

528 1. Thorpe, J., Osei-Owusu, I. A., Avigdor, B. E., Tupler, R. & Pevsner, J. Mosaicism in Human
529 Health and Disease. *Annu. Rev. Genet.* **54**, 1–24 (2020).

530 2. Luquette, L. J. *et al.* Single-cell genome sequencing of human neurons identifies somatic point
531 mutation and indel enrichment in regulatory elements. *Nat Genet* **54**, 1564–1571 (2022).

532 3. Dong, X. *et al.* Accurate identification of single nucleotide variants in whole genome amplified
533 single cells. *Nat Methods* **14**, 491–493 (2017).

534 4. Zhao, Y. *et al.* High-resolution detection of copy number alterations in single cells with
535 HiScanner. *bioRxiv* 2024.04.26.587806 (2024) doi:10.1101/2024.04.26.587806.

536 5. Loh, P.-R. *et al.* Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations.
537 *Nature* **559**, 350–355 (2018).

538 6. Dileep, V. *et al.* Neuronal DNA double-strand breaks lead to genome structural variations and
539 3D genome disruption in neurodegeneration. *Cell* **186**, 4404-4421.e20 (2023).

540 7. Ellegren, H. Heterogeneous mutation processes in human microsatellite DNA sequences. *Nat.*
541 *Genet.* **24**, 400–402 (2000).

542 8. Rajan-Babu, I.-S., Dolzhenko, E., Eberle, M. A. & Friedman, J. M. Sequence composition
543 changes in short tandem repeats: heterogeneity, detection, mechanisms and clinical implications.
544 *Nat. Rev. Genet.* 1–24 (2024) doi:10.1038/s41576-024-00696-z.

545 9. Gymrek, M. *et al.* Abundant contribution of short tandem repeats to gene expression variation in
546 humans. *Nat Genet* **48**, 22–29 (2016).

547 10. Biezuner, T. *et al.* Comparison of seven single cell whole genome amplification commercial
548 kits using targeted sequencing. *Sci. Rep.* **11**, 17171 (2021).

549 11. Willems, T. *et al.* Genome-wide profiling of heritable and de novo STR variations. *Nat Methods*
550 **14**, 590–592 (2017).

551 12. Dolzhenko, E. *et al.* ExpansionHunter: A sequence-graph based tool to analyze variation in short
552 tandem repeat regions. *Bioinformatics* **35**, 4754–4756 (2019).

553 13. Sehgal, A., Ziaei-Jam, H., Shen, A. & Gymrek, M. Genome-wide detection of somatic
554 mosaicism at short tandem repeats. *Bioinformatics* btae485 (2024)
555 doi:10.1093/bioinformatics/btae485.

556 14. Cui, Y. *et al.* A genome-wide spectrum of tandem repeat expansions in 338,963 humans. *Cell*
557 **187**, 6411–6412 (2024).

558 15. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science.
559 *Nat Rev Genet* **17**, 175–188 (2016).

560 16. Niu, B. *et al.* MSIsensor: microsatellite instability detection using paired tumor-normal sequence
561 data. *Bioinformatics* **30**, 1015–1016 (2014).

17. Cortes-Ciriano, I., Lee, S., Park, W.-Y., Kim, T.-M. & Park, P. J. A molecular portrait of microsatellite instability across multiple cancers. *Nat Commun* **8**, 15180 (2017).

18. Park, S. *et al.* Clonal dynamics in early human embryogenesis inferred from somatic mutation. *Nature* **597**, 393–397 (2021).

19. Lodato, M. A. & Walsh, C. A. Genome aging: somatic mutation in the brain links age-related decline with disease and nominates pathogenic mechanisms. *Hum Mol Genet* **28**, R197–R206 (2019).

20. Xia, Y. *et al.* INSIGHT. (2025).

21. Huang, L., Ma, F., Chapman, A., Lu, S. & Xie, X. S. Single-Cell Whole-Genome Amplification and Sequencing: Methodology and Applications. *Annu Rev Genom Hum G* **16**, 1–24 (2015).

22. Gonzalez-Pena, V. *et al.* Accurate genomic variant detection in single cells with primary template-directed amplification. *Proc National Acad Sci* **118**, e2024176118 (2021).

23. Yang, Q. *et al.* PhyloSOLID: robust phylogeny reconstruction from single-cell data despite pervasive errors and extreme sparsity.

24. Yoshida, K. *et al.* Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).

25. Zhang, L. *et al.* Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan. *P Natl Acad Sci Usa* **116**, 9014–9019 (2019).

26. A., L., Michael *et al.* Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018).

27. Zhang, L. *et al.* Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan. *Proc. Natl. Acad. Sci.* **116**, 9014–9019 (2019).

28. Yoshida, K. *et al.* Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).

29. Yang, X. *et al.* Cancer Risks Associated With Germline PALB2 Pathogenic Variants: An International Study of 524 Families. *J. Clin. Oncol.* **38**, 674–685 (2020).

30. Koh, G. C. C. *et al.* A redefined InDel taxonomy provides insights into mutational signatures. *Nat. Genet.* 1–10 (2025) doi:10.1038/s41588-025-02152-y.

31. Gupta, S., Gellert, M. & Yang, W. Mechanism of mismatch recognition revealed by human MutSβ bound to unpaired DNA loops. *Nat. Struct. Mol. Biol.* **19**, 72–78 (2012).

32. Sfeir, A. & Symington, L. S. Microhomology-Mediated End Joining: A Back-up Survival Mechanism or Dedicated Pathway? *Trends Biochem. Sci.* **40**, 701–714 (2015).

33. Scully, R., Panday, A., Elango, R. & Willis, N. A. DNA double-strand break repair-pathway choice in somatic mammalian cells. *Nat. Rev. Mol. Cell Biol.* **20**, 698–714 (2019).

34. Petruseva, I. O., Evdokimov, A. N. & Lavrik, O. I. Molecular Mechanism of Global Genome Nucleotide Excision Repair. *Acta Naturae* **6**, 23–34 (2014).

35. Miller, M. B. *et al.* Somatic genomic changes in single Alzheimer's disease neurons. *Nature* 1–9 (2022) doi:10.1038/s41586-022-04640-1.

36. Fotsing, S. F. *et al.* The impact of short tandem repeat variation on gene expression. *Nat. Genet.* **51**, 1652–1659 (2019).

37. Xiong, X. *et al.* Epigenomic dissection of Alzheimer's disease pinpoints causal variants and reveals epigenome erosion. *Cell* **186**, 4422-4437.e21 (2023).

38. Morabito, S. *et al.* Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. *Nat. Genet.* **53**, 1143–1155 (2021).

39. Hampel, H. *et al.* The Amyloid-β Pathway in Alzheimer's Disease. *Mol. Psychiatry* **26**, 5481–5503 (2021).

40. Xue, Y.-Y. *et al.* CD2AP deficiency aggravates Alzheimer's disease phenotypes and pathology through p38 MAPK activation. *Transl. Neurodegener.* **13**, 64 (2024).

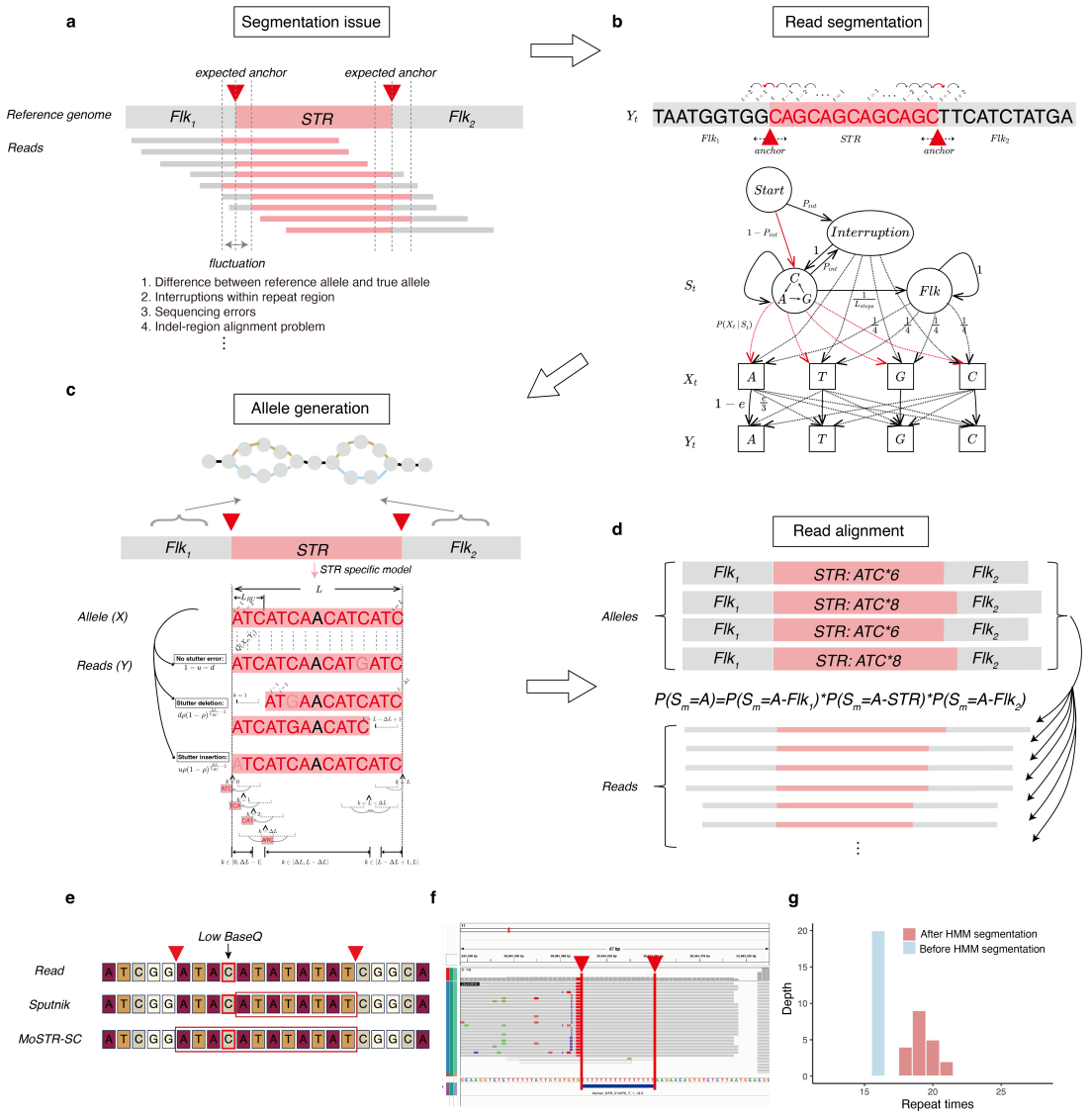
41. Ando, K. *et al.* PICALM and Alzheimer's Disease: An Update and Perspectives. *Cells* **11**, 3994 (2022).

42. Du, Y. *et al.* MKP-1 reduces Aβ generation and alleviates cognitive impairments in Alzheimer's disease models. *Signal Transduct. Target. Ther.* **4**, 58 (2019).

43. Nguyen, C. *et al.* Transcriptional and epigenetic targets of MEF2C in human microglia contribute to cellular functions related to autism risk and age-related disease. *Nat. Immunol.* 1–15 (2025) doi:10.1038/s41590-025-02299-0.

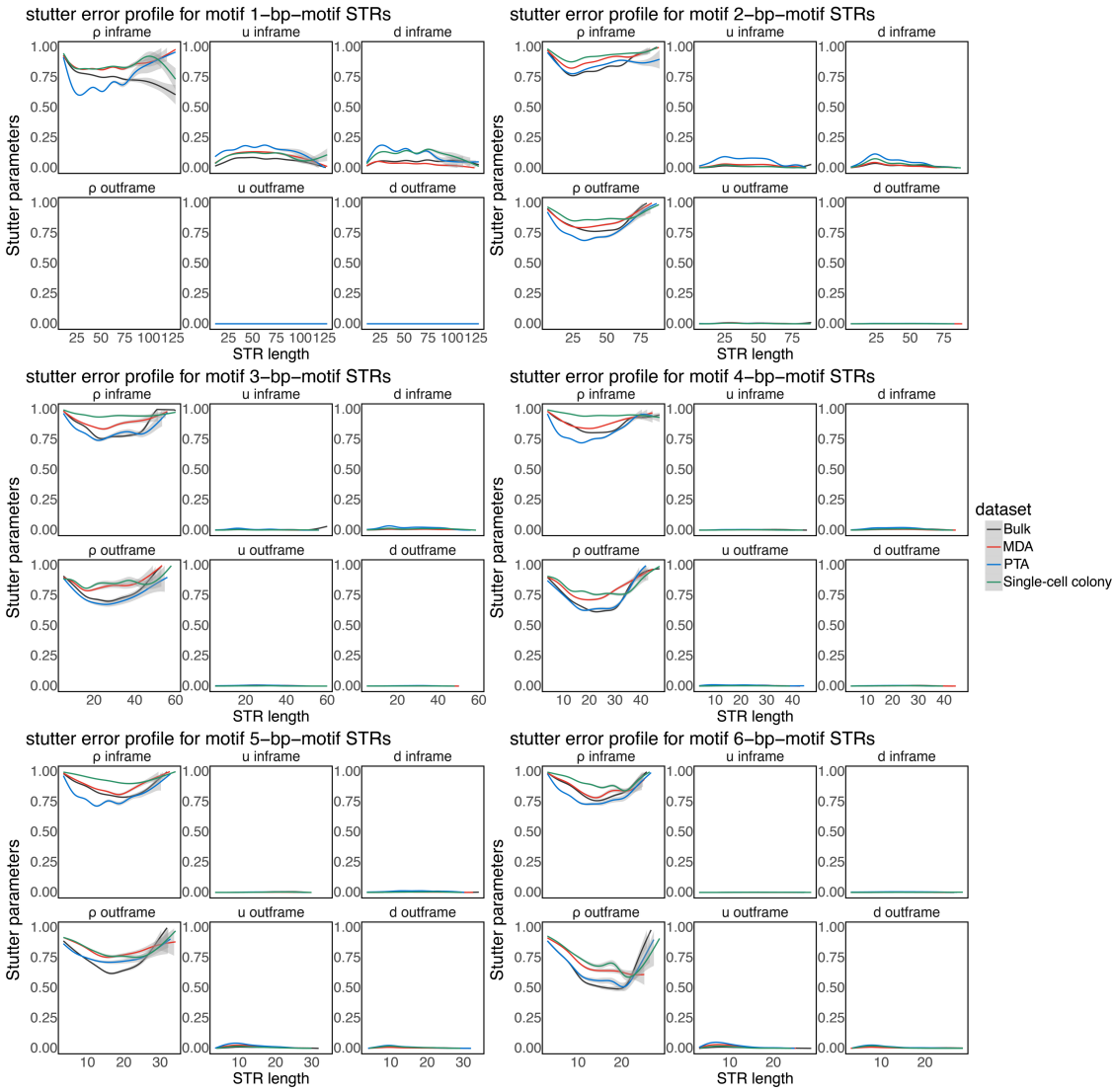
44. Ellegren, H. Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.* **5**, 435–445 (2004).

45. Wang, W., Li, W., Wang, C., Fan, W. & Dou, Y. Accurate detection of mosaic mutations at short tandem repeats from bulk sequencing data. (2026).



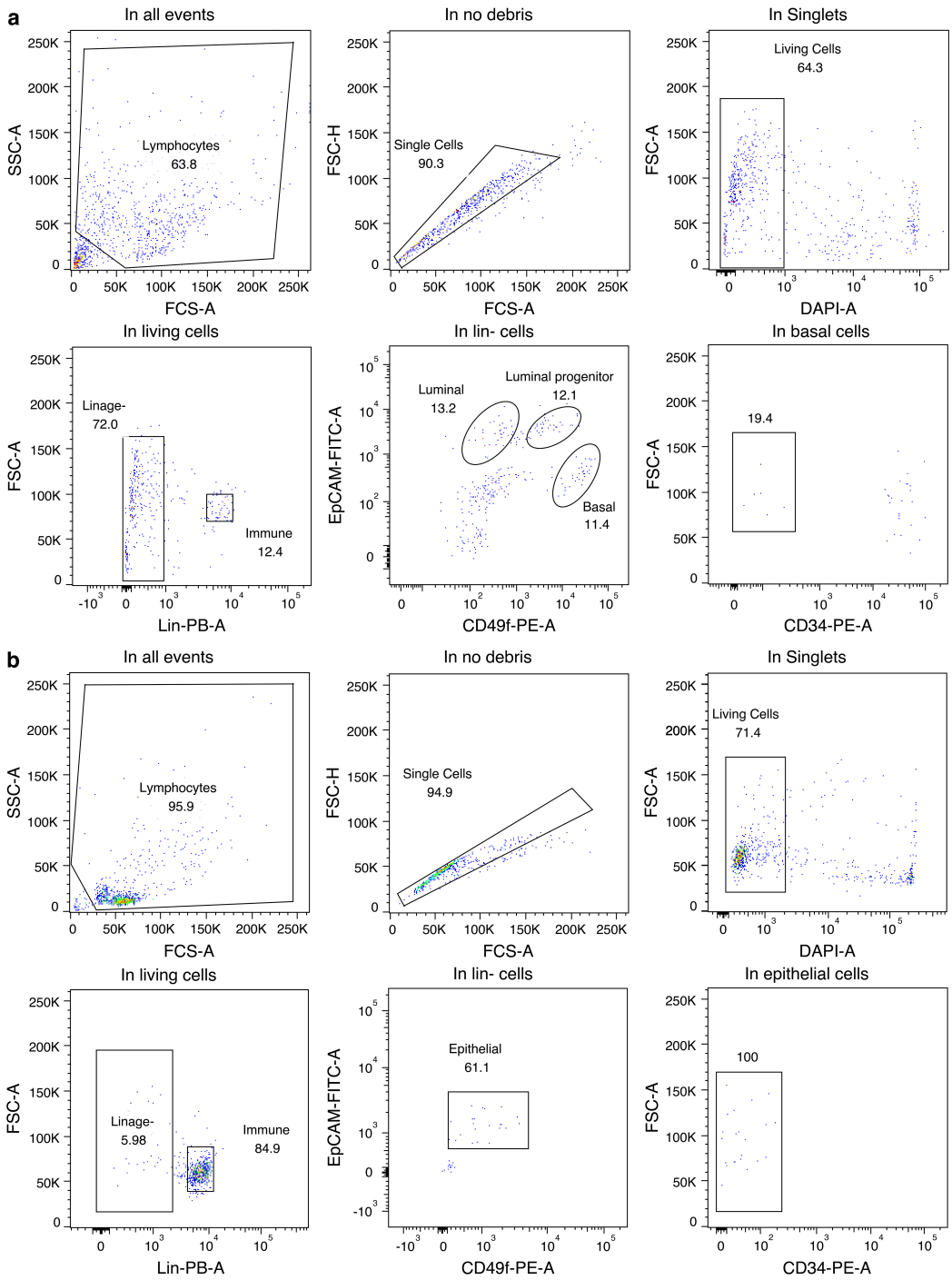
Supplementary Fig. 1 Read segmentation, allele generation and local read alignment based on sequencing reads and information from human reference genome. (a) The boundary ("anchor" in the Fig.) separating the STR-region and flanking regions within sequencing reads often fluctuates due to alignment issues, sequencing errors and other problems. **(b)** An HMM was used for reads segmentation. The three hidden states are "STR", "flanking", and "interruption". Each state exhibits a preference for specific base-pair sequences and emits them with a higher probability. The segmentation boundaries between the STR and two flanking regions were identified by enumerating and selecting the optimal hidden state path. **(c)** STR-specialized local read alignment. Assuming no stutter errors occur, the read alignment likelihood was calculated by multiplying the base agreement probabilities between the read sequences and the hidden allele sequences. In the presence of a stutter error, we assume that DNA slippage occurs only once per read and that the stutter-induced INDEL can occur at any position with equal probability. **(d)** We calculate the probability that each read originates from a specific allele using an STR-specialized local read alignment strategy. **(e-g)** The read segmentation strategy of BayesMonSTR is robust to sequencing errors and can accurately

restore the correct STR length distribution despite indel fluctuations. **(e)** An example demonstrates how the scoring-based Sputnik algorithm, which ignores base quality, leads to segmentation errors. In contrast, BayesMonSTR's HMM-based segmentation strategy incorporates base quality, making it more resilient to sequencing errors. **(f-g)** Compared to BWA alignment, which often struggles with accurately defining STR-flank boundaries, the HMM-based approach of BayesMonSTR effectively tolerates indel fluctuations and restores the correct STR length distribution.



Supplementary Fig. 2 Stutter error estimation across different single-cell sequencing platforms. Stutter-error parameters were derived from a length-based stutter error geometric distribution model, calculated with an EM-algorithm. "Bulk" refers to bulk DNA-seq data, "MDA" refers to single-cell sequencing data amplified with multiple displacement amplification approach, "PTA" refers to single-cell sequencing data amplified with primary template-directed amplification approach, and "SCC" refers to single-cell colony data. STR loci were described by the parameters ρ (rho), u (up), and d (down), which represent the three parameters of the geometric distribution, in-frame and out-frame parameters were

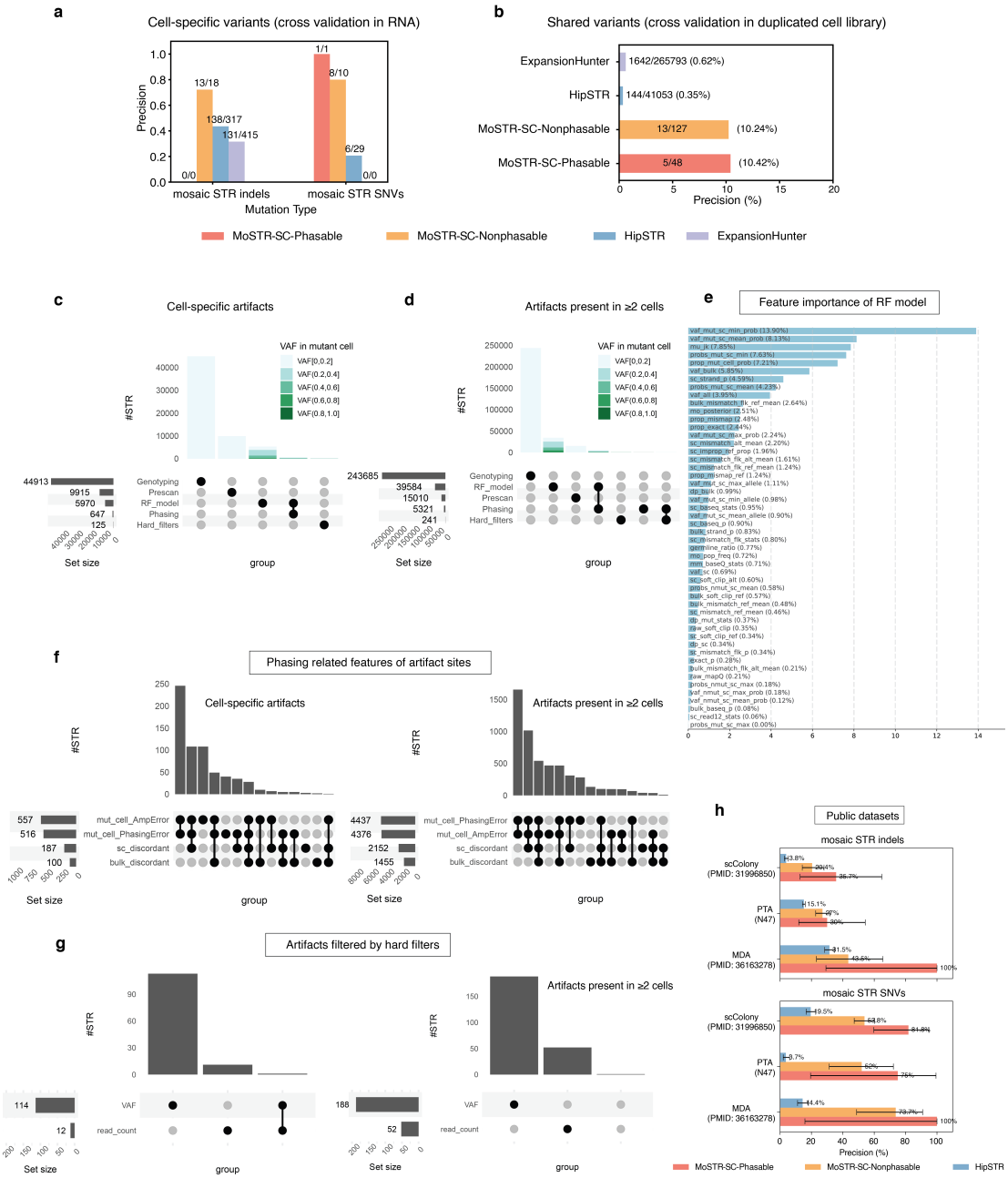
657 estimated separately (Methods). Smoothed trend lines were calculated by the generalized
658 additive model (GAM) with cubic splines, and gray area represents 95% confidence interval.



659

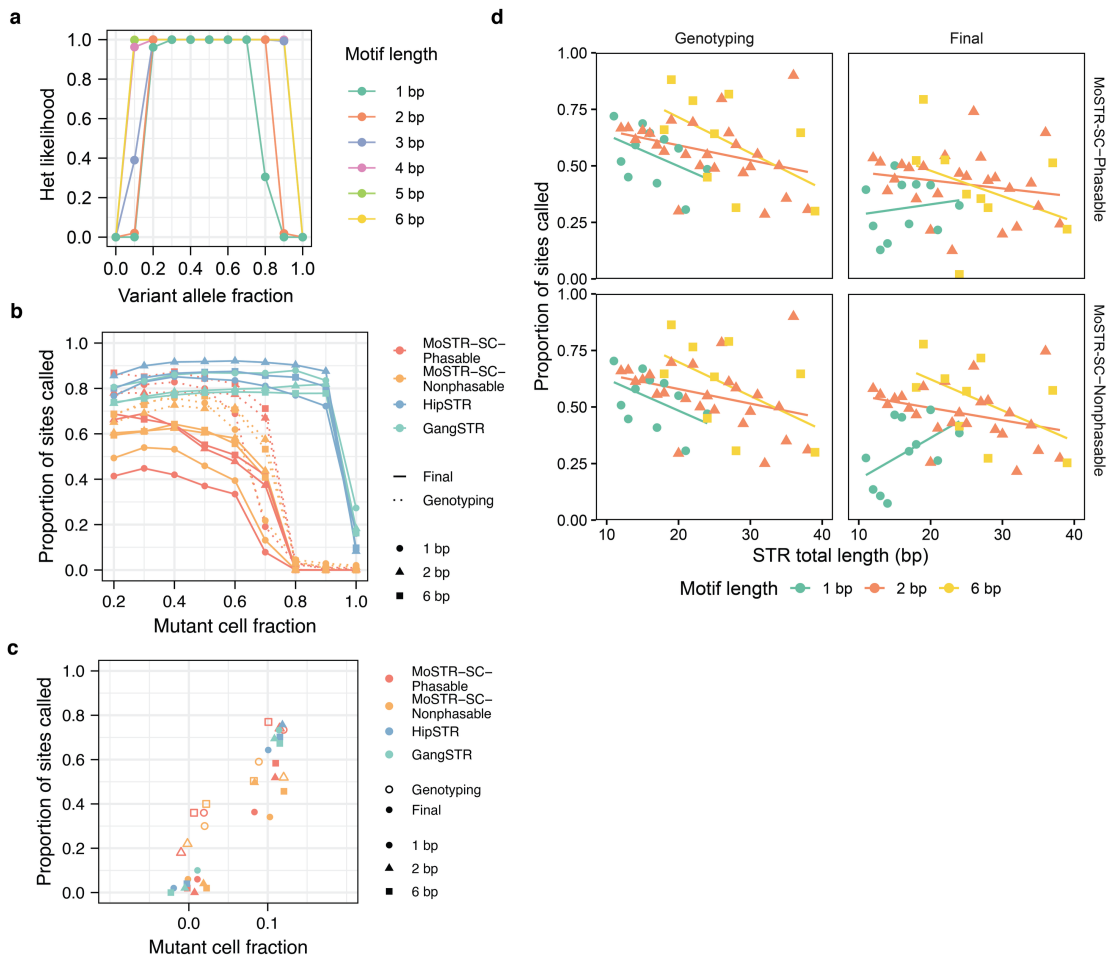
660 **Supplementary Fig. 3 FACS gating strategy for isolating breast-tissue epithelial cells and**
661 **immune cells. (a)** Cell sorting results from paired normal tissue samples. Immune cells were
662 defined as Lin+ cell populations, luminal cells were identified as Lin-EpCAM+CD49f-, luminal
663 progenitor cells were identified as Lin-EpCAM+CD49f+, and basal cells were identified as
664 Lin-EpCAM−CD49f+CD34− cells. **(b)** Cell sorting results from tumor samples. Immune cells
665 were defined as Lin+ cell populations, while epithelial cells were identified as Lin-

EpCAM+CD49f+CD34-. The sorting strategy was used to isolate these specific cell subsets for further analysis.

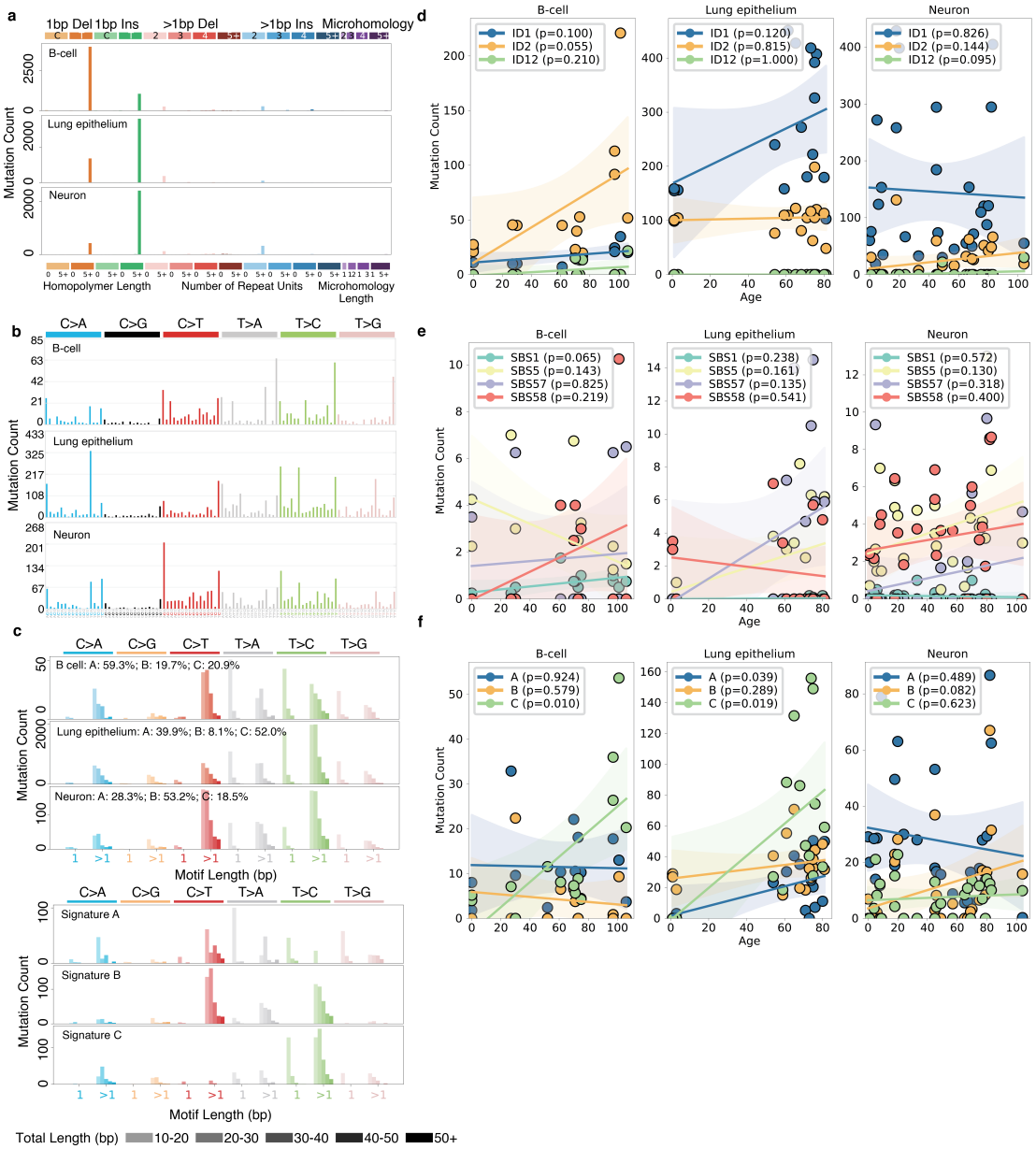


Supplementary Fig. 4 BayesMonSTR outperforms HipSTR and Expansionhunter in detecting mosaic STR mutations from single cells. (a) The precision of cell-specific mutations was evaluated using orthogonal data from single-cell RNA-seq obtained from the same cell (Methods). **(b)** Validation rate of mutations presents in ≥2 cells across different software tools. The sites were extensively evaluated using both orthogonal data and PhyloSOLID (Methods). **(c-d)** Distribution of mutant-cell VAFs for false-positive sites detected by HipSTR and ExpansionHunter. Solid points indicate false positives that were filtered by BayesMonSTR using a specific filtration method. It is evident that BayesMonSTR excluded a majority of these false positives, which had relatively low mutant allele frequencies,

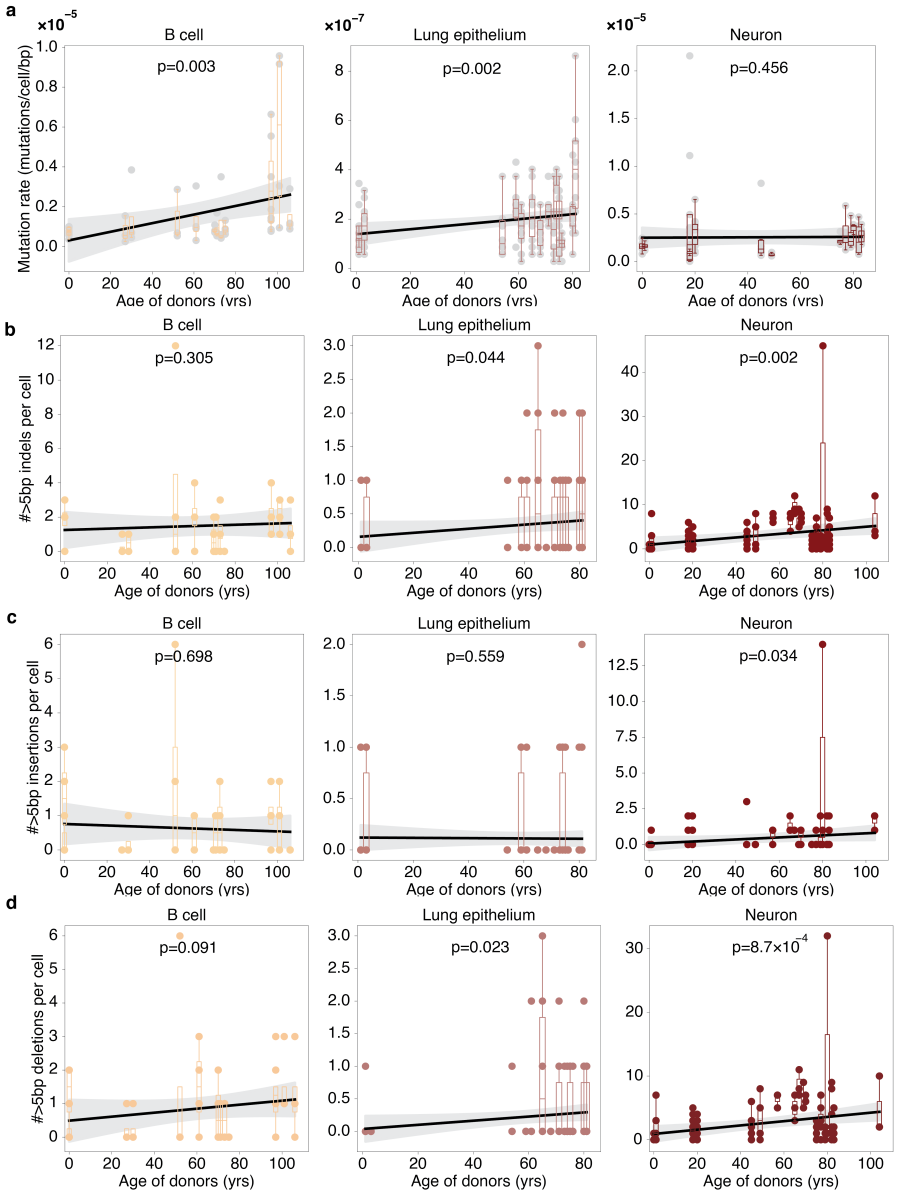
679 during the pre-scan and genotyping stages. Mutations with relatively high mutant-cell VAFs
680 were further excluded by BayesMonSTR using the RF classification model, probabilistic
681 haplotype phasing, and a series of hard filters. **(e)** Expanded analysis highlighting the sites that
682 were excluded by the RF model of BayesMonSTR (Methods). False positives were effectively
683 filtered by BayesMonSTR by using a combination of informative predictors. Explanations of
684 these predictors are provided in Supplementary Table 1. **(f)** Expanded analysis of false-positive
685 sites identified by HipSTR and ExpansionHunter, highlighting key phasing-related features
686 driving the exclusion of false positives by BayesMonSTR. This includes discordant read pair
687 proportion, amplification error-related features. **(g)** Artifact sites excluded by BayesMonSTR
688 using hard filters. **(h)** Performance benchmarking of BayesMonSTR versus HipSTR using
689 additional public datasets. Error bars represent 95% confidence intervals of the validation rates
690 calculated using binomial sampling.



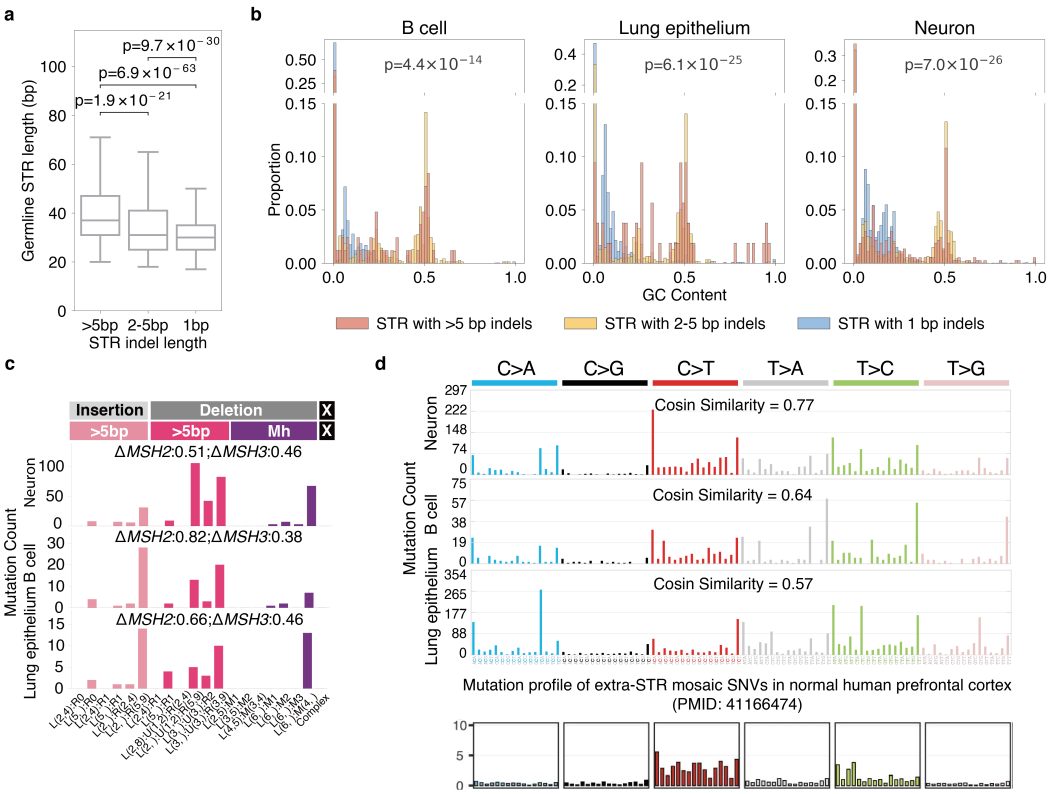
Supplementary Fig. 5 Performance evaluation of BayesMonSTR using simulated data with spike-in mutations. (a) Genotyping likelihoods of germline heterozygous at different variant allele fractions. Reads are assumed with Phred quality score 30 and median stutter error parameters were used for likelihoods calculation, heterozygous likelihoods were normalized along with reference and alternative homozygous likelihoods. **(b-c)** Sensitivity comparison across the two software tools. HipSTR was chosen for the benchmark here, given its superior performance to ExpansionHunter on the real-world data. One-motif mosaic STR insertions within 16-24 bp STR regions were generated using BAMSurgeon. The total number of cells was 10. Dashed lines indicate the raw mutation callings after genotyping of BayesMonSTR, while solid lines indicate final call sets. B illustrates the sensitivity detecting shared mutations and C illustrate cell-specific mutations. **(d)** Sensitivity variation across different STR total lengths. The total cell number was 10 and mutations spiked in were 1-motif insertions.



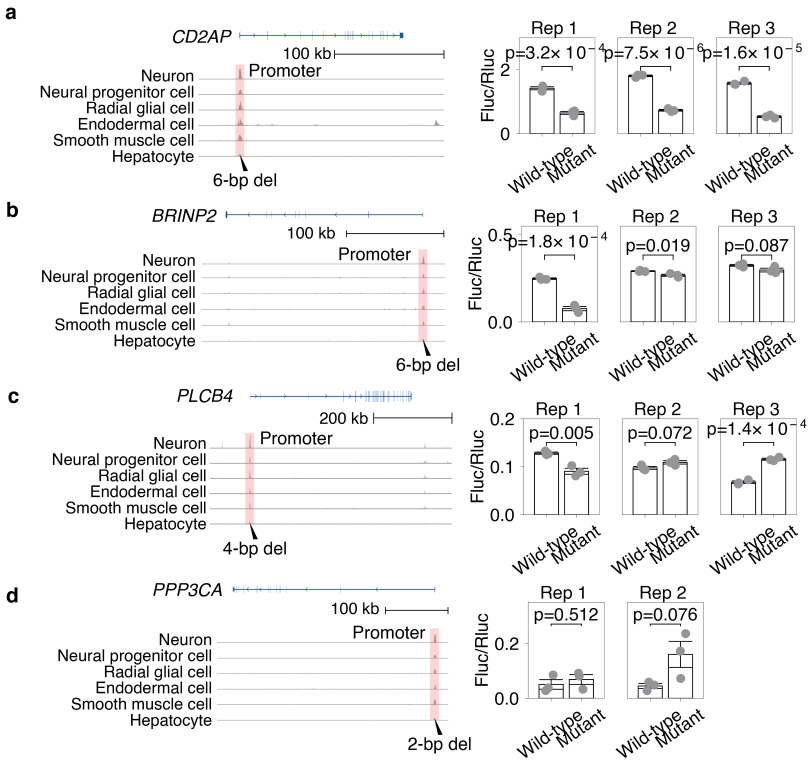
Supplementary Fig. 6 Mutation signature analysis. Cell type-level mutation profiles of mosaic STR-indels **(a)** and mosaic STR-SNVs **(b-c)** plotted with SigProfiler. Specifically, the mutation signature of STR-SNVs in panel b was defined by us, taking into account the nucleotide context of the entire STR region; while panel c represents the classical 96-cat mutation signature, which considers only the 3-nucleotide context. **(d-f)** Accumulation of mosaic STR mutations derived from different mutation signatures by age. The mutation burden associated with each signature was quantified for every individual (Method). From top to end, mutation signatures were extracted from STR-indels by SigProfiler, STR-SNVs by custom method, and STR-SNVs by SigProfiler, respectively. Each data point represents the mutation burden of an individual, with the solid line representing the fitted curve relating mutation burden to donor age.



Supplementary Fig. 7 STR indels across cell types. (a) 1-bp STR indels accumulate with age in B cells and lung epithelial cells. In contrast, the mutation rate remains relatively stable in non-dividing neurons across donors of different ages. **(b–d)** Relationship between age and the total burden per cell of STR indels >5 bp (b), STR insertions >5 bp (c), and STR deletions >5 bp (d) across three cell types. In all three cell types, the burden of STR deletions >5 bp shows a positive correlation with age. However, only in neurons does the burden of indels, insertions, and deletions consistently increase with age. Linear regression equations were derived using the least squares method. P-values were calculated using a one-tailed t-test to test whether the regression coefficient is significantly greater than zero. The shaded areas around each curve represent the 95% CIs, assuming the differences between predicted values and observed values follow a t-distribution.



Supplementary Fig. 8 Mutational mechanism analysis of mosaic STR mutations. (a) Longer indels preferentially occur in STRs that are themselves longer. P-value was calculated with a two-tailed Mann-Whitney U test. **(b)** STRs with longer indels exhibit significantly higher GC content compared to STRs with 1-bp indels. The P-value was calculated with a two-tailed Kolmogorov-Smirnov (KS) test and reflects the comparison between STRs with indels longer than 5 bp and those with 1-bp indels. **(c)** STR indels greater than 5bp have mutation profiles similar to those caused by the *MSH2* and *MSH3* knockouts. The x-axis represents the indel types defined by a redefined InDel taxonomy, where L stands for the indel length, R represents the repeat length of the repetitive motif in the 3′ sequence, and M indicates the microhomology length. Mh refers to microhomology. The mutation signature similarity was computed via cosine similarity. **(d)** STR-region SNVs largely mirrored the established patterns of extra-STR mosaic SNVs in normal human prefrontal cortex from PMID 41166474. The mutation signature similarity was computed via cosine similarity.



Supplementary Fig. 9 Luciferase assay results of somatic STR indels in regulatory regions. Normalized levels of H3K4me3 and H3K27ac—key histone modifications linked to promoter and enhancer activities across cell types were obtained from the ENCODE project. Pink regions indicate the promoter or enhancer regions of the gene, with indels highlighted by black triangles. Luciferase assays in HEK293T cells revealed that a 6-bp deletion in the TSS of *CD2AP* and a 6-bp deletion in the TSS of *BRINP2* significantly disrupted gene expression ($P < 0.05$, two-tailed t-test). Furthermore, a 4-bp deletion in the TSS of *PLCB4* and a 2-bp deletion in the TSS of *PPP3CA* enhanced gene expression with marginal significance ($P < 0.1$, two-tailed t-test). Error bars represent standard deviation (3 biological replicates).

Methods

Sequence alignment and data preprocessing

Raw FASTQ reads were aligned to the GRCh37 reference genome (human_g1k_v37_decoy) using BWA-MEM (v0.7.18). Duplicate reads were marked with Picard MarkDuplicates (v2.27), and base quality score recalibration (BQSR) was performed with the Genome Analysis Toolkit (GATK v4.2). Germline variants were called with GATK HaplotypeCaller on both MDA single-cell and matched bulk data and filtered using Variant Quality Score Recalibration (VQSR) following GATK Best Practices. Common germline heterozygous SNPs (gnomAD[1] allele frequency >0.1%) were selected for downstream haplotype phasing with Eagle2 (v2.4.1)[2].

Stutter-error estimation

To reduce amplification- and sequencing-related artifacts introduced by *in vitro* polymerase slippage, BayesMonSTR models STR stutter errors using a geometric-distribution[3]. The stutter error probability $p(r_{m,c}|a=j)$ for m-th read r for cell c in the given allele j was modeled as follows:

$$P(r_{m,c}|a=j) = \begin{cases} 1-u-d, & r_{m,c}=r_a \\ u\rho(1-\rho)^{r_{m,c}-r_a-1}, & r_{m,c}>r_a \\ d\rho(1-\rho)^{r_a-r_{m,c}-1}, & r_{m,c}<r_a \end{cases} \quad (1)$$

Where:

- u denotes the total mutation rate of artifact insertions.
- d denotes the total mutation rate of artifact deletions.
- ρ denotes the success probability of the geometric model, which belongs to the interval (0,1].

Assuming independence across reads, the genotyping likelihood for each diploid cell is computed using equation 2:

$$P(D_c|G_c=(j,k),\theta_{stutter,cell}) = \prod_{m=1}^{n_{reads,c}} \sum_{a\in\{j,k\}} \frac{1}{2} \begin{cases} 1-u-d & r_{m,c}=r_a \\ u\rho(1-\rho)^{r_{m,c}-r_a-1} & r_{m,c}>r_a \\ d\rho(1-\rho)^{r_a-r_{m,c}-1} & r_{m,c}<r_a \end{cases} \quad (2)$$

Where:

- G_c denotes the genotype of a given cell.
- j and k denote the two haplotype alleles.
- D_c denotes the sequencing reads of a given cell c.
- The parameter set θ denotes the stutter model parameters (u, d, ρ), as described in formular (1).

Of note, reads with mapping quality < 20, as well as reads with mean base quality < 20, were excluded before stutter-parameter estimation.

Parameter initialization:

- Given n candidate alleles, the initial population frequency for allele a, f_a, was set to $\frac{1}{n}$, representing equal prior probability across all alleles.

- The initial mosaic allele fraction parameter μ_{msi} was set to 0.001.
- Stutter error parameters u, d and ρ, which define the geometric model, were initialized to 0.01, 0.01, and 0.9, respectively.

E-step:

In the E-step, posterior probabilities of individual germline genotypes were computed using equation 3:

$$\begin{aligned}
&P(G_i^t = (j,k)|G_c^t, D_c, D_{bulk}, \theta^t_{stutter,cell}, \theta^t_{stutter,bulk}) = \\
&P(G_i^t = (j,k)) * P(D_{bulk}|G_{bulk}^t = (j,k), \theta^t_{stutter,bulk}) * \prod_{c=1}^{n_{cells,i}} [P(G_c^t|G_i^t) * P(D_c|G_c^t, \theta^t_{stutter,cell})] = \\
&2f_j^t f_k^t * \prod_{m=1}^{n_{reads,i,bulk}} \frac{1}{2} \sum_{a \in j,k} P(S_{m,i} = a|\theta^t_{stutter,bulk}) * \prod_{c=1}^{n_{cells,i}} \sum \begin{cases} \frac{1}{2}\mu_{msi} * \sum_{-k \in A_i \setminus \{k\}} P(D_c|G_c^t = (j,-k), \theta^t_{stutter,cell}) \\ (1-\mu_{msi}) * P(D_c|G_c^t = (j,k), \theta^t_{stutter,cell}) \\ \frac{1}{2}\mu_{msi} * \sum_{-j \in A_i \setminus \{j\}} P(D_c|G_c^t = (-j,k), \theta^t_{stutter,cell}) \end{cases}
\end{aligned} \tag{3}$$

Where:

- f_j denotes population allele frequency for allele j.
- G_i denotes the individual germline genotype for individual i, and G_c denotes the cell genotype.
- θ denote the set of stutter error parameters as previously described.
- D_c denotes the sequencing reads of a given cell c.

Here we assume G_i is the same as G_{bulk}, and the prior probabilities of individual genotypes were calculated based on Hardy-Weinberg Theorem. When j equals to k, the prior probability of a homozygous germline genotype is given by f_j^2.

M-step:

We then estimated the error parameters u, d and ρ, and the length-based population allele frequency f at $t+1$-th step for stutter errors in the M-step, and were calculated by equations 4-7, as follows:

$$u^{t+1} = \frac{\sum_{c=1}^{N_{ind}} \sum_{m=1}^{N_{reads,c}} P\left(S_{m,c,i}, r_{m,c,i}|\theta^t_{stutter}\right) I\left(r_{m,c,i} > s_{m,c,i}\right)}{N_{reads}} \tag{4}$$

$$d^{t+1} = \frac{\sum_{c=1}^{N_{ind}} \sum_{m=1}^{N_{reads,c}} P\left(S_{m,c,i}, r_{m,c,i}|\theta^t_{stutter}\right) I\left(r_{m,c,i} < s_{m,c,i}\right)}{N_{reads}} \tag{5}$$

$$\rho^{t+1} = \frac{\sum_{c=1}^{N_{ind}} \sum_{m=1}^{N_{reads,c}} P\left(S_{m,c,i}, r_{m,c,i}|\theta^t_{stutter}\right) I\left(r_{m,c,i} \neq s_{m,c,i}\right)}{\sum_{c=1}^{N_{ind}} \sum_{m=1}^{N_{reads,c}} P\left(S_{m,c,i}, r_{m,c,i}|\theta^t_{stutter}\right) \left|r_{m,c,i} - s_{m,c,i}\right|} \tag{6}$$

$$f_j^{t+1} = \frac{1}{2N_{ind}} \sum_{i=1}^{N_{ind}} \sum_{k=1}^{A_i} \left(P\left(G_i^t = (j,k)\middle|G_{c,i}^t, D_{c,i}, D_{bulk,i}, \theta^t_{stutter}\right) + P\left(G_i^t = (k,j)\middle|G_{c,i}^t, D_{c,i}, D_{bulk,i}, \theta^t_{stutter}\right)\right) \tag{7}$$

Where:

- N_{ind} denotes total number of individuals.

- The indicator function I is used to restrict the computation to reads with insertion, deletion and nonzero stutter for u, d and ρ respectively.
- $p(S, r|\theta)$ is read assignment probability for read $r_{m,c}$ given allele pool A_i (see formula 8).
- G_i denotes the individual germline genotype for individual i, and G_c denotes the cell genotype.
- θ denote the set of stutter error parameters as previously described.
- D_c denotes the sequencing reads of a given cell c.
- N_{reads} denotes the total number of samples with the same sequencing method. (including bulk, WGA, single cell colonies, etc).
- $N_{reads,c}$ denotes the total number of reads for sample c.
- the indicator function I is used to restrict the computation to reads with insertion.
- deletion and nonzero stutter for u, d and ρ respectively.
- θ denoted these stutter parameters.
- f_j denotes population allele frequency for allele j.

$p(S, r|\theta)$ is read assignment probability for read $r_{m,c}$ given allele pool A_i:

$$\begin{aligned}
P(S_{m,c,i}, r_{m,c,i}|\theta^t_{stutter,cell}) &= \sum_{a\in\{A_i\}} P(S_{m,c,i}=a) * P(r_{m,c,i}|S_{m,c,i}=a, \theta^t_{stutter,cell}) = \\
&= \sum_{a\in\{j,k\}} \sum_{j\in\{A_i\},k\in\{A_i\}} P(S_{m,c,i}=a|G^t_{c,i}=(j,k)) * \\
&\quad P(G^t_{c,i}=(j,k)) * P(r_{m,c,i}|S_{m,c,i}=a, \theta^t_{stutter,cell}) = \\
&= \sum_{a\in\{j,k\}} \sum_{j\in\{A_i\},k\in\{A_i\}} \sum_{G^t_i} P(S_{m,c,i}=a|G^t_{c,i}=(j,k)) * \\
&\quad P(G^t_{c,i}=(j,k), G^t_i|D_{c,i}, \theta^t_{stutter,cell}) * P(r_{m,c,i}|S_{m,c,i}=a, \theta^t_{stutter,cell}) = \\
&= \sum_{a\in\{j,k\}} \sum_{j\in\{A_i\},k\in\{A_i\}} \sum_{G^t_i} P(S_{m,c,i}=a|G^t_{c,i}=(j,k)) * \\
&\quad P(G^t_{c,i}=(j,k)|G^t_i) * P(D_{c,i}|G^t_{c,i}=(j,k), \theta^t_{stutter,cell}) * \\
&\quad P(G^t_i) * P(r_{m,c,i}|S_{m,c,i}=a, \theta^t_{stutter,cell})
\end{aligned} \tag{8}$$

Where:

- a is one of the alleles in the alleles pool $A_i = \{j, k\}$ for individual i.
- other parameters are defined in Equation (3)

Probabilistic read segmentation

Accurate detection of repeat regions within sequencing reads is essential for generating candidate STR alleles. Traditional segmentation methods, which rely on alignment coordinates and repeat patterns, often suffer from inaccuracies due to mapping biases and interruptions. These issues lead to fluctuations in the 'anchor points' used to distinguish repeat and flanking regions, causing the anchor points to deviate from their expected coordinate across different reads (Extended Data Fig. 1a). To address these

issues, we utilized a Hidden Markov Model (HMM) approach to identify repeat regions in sequencing reads.

Notably, our HMM model integrates information from both sequencing reads (including sequencing data and quality scores) and the reference genome (repeat-region and flanking region sequences). It also incorporates prior knowledge of in-frame and out-of-frame interruptions derived from the human reference genome. This design enables robustness to sequence interruptions, sequencing errors, and mutations within STR regions, while integrating base quality scores directly into the segmentation process (Extended Data Fig. 1).

Specifically, the HMMs were designed to model the STR sequences using three hidden blocks (Extended Data Fig. 1b):

A. STR block: Represents perfect repeat motifs, capturing the repetitive structure at each STR locus. For example, the motif CAG is modeled with three hidden states—S_C, S_A, and S_G —within this block.
B. Interruption block: Models mismatches and indels within repeat regions. Similar to profile HMM[4], this block accounts for variations inside the repeat tract.
C. Flanking block: Represents sequences flanking the STR. In addition to a general flanking state with equal emission probabilities for A, C, G and T, we introduce a locus-specific 3-bp known flanking states extracted from the human reference genome. These states incorporate sequence context to improve segmentation accuracy.

To initialize the transition probabilities, the initial transition probability from the STR block to the flanking block was defined as 1/L, where L corresponds to the expected segmentation distance before entering the flanking region. For homopolymers, L was set to 5 bp, while for non-homopolymers, L was set to twice the motif length. Furthermore, the initial transition probability from the STR block to the interruption block is estimated through an extensive grid search using a reference panel[3].

The emission probability matrix integrates both the expected STR motif sequence and base quality scores to improve robustness against sequencing errors. Each hidden state corresponds to an expected nucleotide that receives the highest emission probability. In the flanking block, known 3-bp flanking sequences from the reference genome are assigned predefined emission probabilities, while nucleotides outside this region have equal probabilities across A, C, G, and T. For a given hidden state, if the expected base is X_t, the observed base is Y_t, and the sequencing error rate is e, the emission probability for this base is calculated as follows:

$$P(Y_t \mid X_t) = \begin{cases} 1-e, & Y_t = X_t \\ \frac{e}{3}, & Y_t \neq X_t \end{cases} \tag{9}$$

Where:

- Y_t indicates the observed base in sequencing reads.
- X_t indicates the expected base given a specific hidden state.
- e indicates sequencing error rate.

Finally, the boundaries between STR and flanking regions are determined by finding the maximum-likelihood hidden state path using the Viterbi algorithm. Based on this optimal state path, sequencing reads are segmented into STR and flanking segments. In cases of large deletions where the mutant allele length falls below the predefined step length—leading to potential segmentation failure—a BWA-based coordinate mapping approach is used as a backup. Similarly, due to the high sequence complexity at imperfect STR loci with 4-bp, 5-bp, or 6-bp motifs containing interruptions, segmentation performance may be compromised; for such loci, the BWA-based coordinate method is also applied.

STR-specific read alignment

STR sequences from the segmented reads are used to generate candidate alleles, each comprising the core STR region and 5 bp of flanking sequence on both sides. Segmented reads are subsequently aligned to the candidate alleles using a STR-specific alignment method[3] to calculate the alignment likelihood. This method accounts for both potential PCR stutter errors and sequencing errors. Specifically, three possible scenarios are considered for the sequencing reads: no stutter error, stutter deletion, and stutter insertion. The calculation of alignment likelihood varies slightly for each scenario, as illustrated in Extended Data Fig. 1c.

Given a haplotype X of length L, if read Y is assumed to be free of stutter errors, the likelihood of the observed alignment is determined by the agreement between each base i in the read and its corresponding base in the haplotype:

$$P(X,Y) = (1-u-d) * \prod_{i=1}^{L} Q(X_i, Y_i) \tag{10}$$

Where:

- X denotes the given STR haplotype.
- Y denotes the STR haplotype observed in the read.
- L denotes the STR length of X haplotype.
- u denotes stutter insertion rate from the estimated stutter model.
- d denotes stutter deletion rate from the estimated stutter model.
- $Q(X_i, Y_i)$ represents the agreement score between the i-th base of haplotype X and the i-th base of haplotype Y,as defined in Equation (9).

Otherwise, if read Y contains a stutter deletion of haplotype X—with a deletion size of ΔL—we assume that the deletion could occur at any position k within the haplotype, where k ranges from 1 to total_haplotype_length $-\Delta L + 1$. The read alignment likelihood is then determined jointly by the probability of stutter deletion and the base-wise agreement between the read and the haplotype. The overall likelihood is obtained by averaging over all possible deletion positions:

$$P(X,Y) = d\rho(1-\rho)^{\frac{\Delta L}{L_{RU}}-1} * \frac{1}{L-\Delta L+1} \sum_{k=1}^{L-\Delta L+1} \prod_{i=1}^{k-1} Q(X_i, Y_i) \prod_{i=k+\Delta L}^{L} Q(X_i, Y_{i-\Delta L}) \quad (11)$$

Where:

- ΔL denotes deletion size in read Y.
- L_{RU} denotes the motif length (in bp).
- ρ denotes the success probability of the geometric model, which belongs to the interval (0,1].
- k denotes the deletion position in the haplotype X.
- Other parameters are defined in Equation (10).

In the third scenario, where the read Y results from a stutter insertion in haplotype X (insertion size ΔL), we assume the inserted sequence arises from a local duplication of adjacent sequence. Consequently, the duplication can originate from either the left or right flank of the insertion site. Due to boundary constraints—insertions within the first ΔL positions can only be duplicated from the right flank, and those within the last ΔL positions only from the left—only one duplication direction is considered for such edge cases. The alignment likelihood is determined jointly by the stutter insertion rate and the base-wise agreement between the read and the haplotype. As before, the overall likelihood is obtained by averaging over all possible insertion positions and duplication directions:

$$P(X,Y) = \left(u*\rho*(1-\rho)^{\frac{\Delta L}{L_{RU}}-1}*\frac{1}{(L+1)}\right)*\sum\begin{cases} \sum_{k=0}^{\Delta L-1}\left(\prod_{i=1}^{k+\Delta L} Q(X_i,Y_i)\right)*\left(\prod_{i=k+1}^{L} Q(X_i,Y_{i+\Delta L})\right) \\ \sum_{k=\Delta L}^{L-\Delta L}\left(\prod_{i=1}^{k} Q(X_i,Y_i)\right)*\left[\frac{\prod_{i=k+1}^{k+\Delta L} Q(X_i,Y_i) * \prod_{i=k+1}^{k+\Delta L} Q(X_{i-\Delta L},Y_i)}{2}\right]*\left(\prod_{i=k+1}^{L} Q(X_i,Y_{i+\Delta L})\right) \\ \sum_{k=L-\Delta L+1}^{L}\left(\prod_{i=1}^{k} Q(X_i,Y_i)\right)*\left(\prod_{i=k+1}^{k+\Delta L} Q(X_{i-\Delta L},Y_i)\right)*\left(\prod_{i=k+1}^{L} Q(X_i,Y_{i-\Delta L})\right) \end{cases} \quad (12)$$

Where:

- ΔL denotes the insertion size in read Y.
- L_{RU} denotes the motif length.
- ρ denotes the success probability of the geometric model, which belongs to the interval (0,1].
- k denotes the insertion position in the haplotype X.
- Other parameters are defined in Equation (10).

We use alignment likelihoods to estimate the probability that each read originates from a given haplotype, thereby determining the most likely haplotype of origin for each read (Extended Data Fig. 1d). The mutation type (insertion, deletion, or mismatch) was further determined by pairwise alignment of the inferred mutant haplotype to the germline haplotype using pairwise aligner[5] (v1.81). This refinement alignment incorporated 4-bp flanking sequences on each side.

Estimating global whole-genome amplification bias

To estimate whole-genome allelic balance for each single cell, we modeled alternative allele counts using a beta-binomial distribution. Germline heterozygous SNPs (hSNPs)

were defined as sites with matched-bulk VAF from 0.4 to 0.6 and present in dbSNP. The two shape parameters of the Beta distribution, α and β were estimated as follows:

$$\alpha = \frac{dp * m_1 - m_2}{dp * (\frac{m_2}{m_1} - m_1 - 1) + m_1} \tag{13}$$

$$\beta = \frac{(dp - m_1) * (dp - \frac{m_2}{m_1})}{dp * (\frac{m_2}{m_1} - m_1 - 1) + m_1} \tag{14}$$

where:

- $m_1 = mean(n_{alt})$.
- $m_2 = \frac{\sum_{m=1}^{n_{site}} n_{alt,m}^2}{n_{site}}$.
- dp is defined as the median depth of hSNP loci.

Notably, the hSNPs collected for calculating allele imbalance parameters were all of the same median depth, such as 20x. n_{alt} refers to the alternative read count, and n_{site} refers to the total number of hSNP loci.

Estimating local amplification bias using a piecewise gaussian process regression

In whole-genome amplification, uneven amplification across DNA fragments creates locally correlated allelic bias. We addressed this by segmenting reads into fragments using an HMM and then applying piecewise Gaussian process regression to quantify bias within each segment.

To model the correlation of allelic balance, we implemented a piecewise Gaussian process regression framework. First, reads within individual cells were partitioned into independent segments using a hidden Markov model (HMM). Within each segment, the haploid allelic fraction (AF_h) values of the same haplotype h, derived from phased hSNPs (phased with Eagle2 as described previously), along with their genomic coordinates, served as inputs for the regression. The Gaussian process is formally defined as:

$$B_h = f(x) \sim GP(\mu = 0, K(x, x'))$$
$$K(x, x') = \sigma_{frag}^2 \exp(-\frac{(x - x')^2}{2l_{frag}^2}) \tag{15}$$

and the allele fraction is represented using a logistic transform:

$$AF_h = \frac{1}{1 + e^{-B_h}}, \quad B_h \in (-\infty, +\infty) \tag{16}$$

Where:

- allele imbalance B_h given hSNPs is modelled by Gaussian process GP with zero mean function.
- the radial basis function $K(x, x')$ with parameter σ and l accounting for fragments attribution.

To prevent discontinuities in the regression line caused by allele dropout, the haploid allelic fraction (AF_h) was constrained to the interval (10^{-7}, $1 - 10^{-7}$). This model was

subsequently applied to predict AF values at each candidate STR loci. Genomic regions containing phased SNPs were excluded if their correlation curves showed sharp fluctuations, specifically those with a mean squared error (MSE) exceeding 0.5, to filter out intervals with potential phasing errors.

Calculation of mosaic genotyping likelihood with bulk sequencing data

We made no specific assumption about the allele frequencies of mosaic mutations; instead, we adopted a uniform prior distribution over the interval [0, 1]. Based on this, the mosaic genotyping likelihood given bulk sequencing data can be calculated with the following formula:

$$P(D|G_i = mosaic) = \int_0^1 (\theta)^r (1-\theta)^{depth-r} d\theta = \binom{depth}{r} \beta(r+1, depth-r+1) \qquad (17)$$

Where:

- r indicates alternative read count in the bulk sequencing data.
- $depth$ indicates total read depth in the bulk seq data.
- θ indicates the prior distribution of variant allele frequency of mosaic mutations.

Joint Genotyping

Based on population allele frequencies, sequencing reads and stutter error profiles, BayesMonSTR calculated genotyping posteriors of candidate mosaic STR mutations with a haplotype-based EM-algorithm.

Parameter initialization:

Population frequencies f_a were set to be with equal probability. Haplotype phasing probability $P(j-h_1, k-h_2|G_i)$ was set to 0.5. μ_{msi} was set to 0.001. Stutter error parameters u, d and ρ for each site were calculated with previous procedures.

E-step:

BayesMonSTR employed a hierarchical Bayesian network to incorporate population, bulk and single cell-level data. We utilized an EM algorithm to calculate maximum likelihood of latent parameters.

Similar to the length-based EM procedure, the E-step for estimating individual germline genotypes at sequence-level resolution was performed as follows:

$$\begin{aligned} &P(G_i^t = (j,k)|D_c, D_{bulk}) = P(G_i^t = (j,k)) * P(D_{bulk}|G_{bulk}^t = (j,k)) * \prod_{c=1}^{n_{cells,i}} \sum_{G_c^t} [P(G_c^t|G_i^t = (j,k)) * P(D_c|G_c^t)] = \\ &2f_j^t f_k^t * \prod_{m=1}^{n_{reads,bulk}} \sum_{a\in\{j,k\}} \frac{1}{2} P(r_{m,i}, Q_{m,i}|S_{m,i} = a) * \prod_{c=1}^{n_{cells,i}} \sum \begin{cases} \sum_{\neg k\in A_i\setminus\{k\}} \mu_{i,k\to\neg k}^t * (1-\sum_{\neg j\in A_i\setminus\{j\}} \mu_{i,j\to\neg j}^t) * P(D_c|G_c^t = (j,\neg k)) \\ (1-\sum_{\neg k\in A_i\setminus\{k\}} \mu_{i,k\to\neg k}^t) * (1-\sum_{\neg j\in A_i\setminus\{j\}} \mu_{i,j\to\neg j}^t) * P(D_c|G_c^t = (j,k)) \\ \sum_{\neg j\in A_i\setminus\{j\}} \mu_{i,j\to\neg j}^t * (1-\sum_{\neg k\in A_i\setminus\{k\}} \mu_{i,k\to\neg k}^t) * P(D_c|G_c^t = (\neg j,k)) \end{cases} \end{aligned} \qquad (18)$$

If the germline genotype is homozygous, the prior probability of G_i is f_j^2 instead of $2f_jf_k$.

Equation 2 and 19 were used for unphased and phased genotype likelihood calculation, respectively.

$$
\begin{aligned}
P(D_c|G_c^t=(j,k)) &= \prod_{m=1}^{n_{reads,c}} \sum \begin{cases} P(O_{m,c}|S_{m,c}=j, S_{m,c}=h_1) * P_{j-h_1}^t \\ P(O_{m,c}|S_{m,c}=j, S_{m,c}=h_2) * P_{j-h_2}^t \\ P(O_{m,c}|S_{m,c}=k, S_{m,c}=h_1) * P_{j-h_2}^t \\ P(O_{m,c}|S_{m,c}=k, S_{m,c}=h_2) * P_{j-h_1}^t \end{cases} \\
&= \prod_{m=1}^{n_{reads,c}} \sum \begin{cases} AF_{c,j-h_1} * P(r_{m,c}|a=j) * P_{j-h_1}^t \\ AF_{c,j-h_2} * P(r_{m,c}|a=j) * P_{j-h_2}^t \\ AF_{c,k-h_1} * P(r_{m,c}|a=k) * P_{j-h_2}^t \\ AF_{c,k-h_2} * P(r_{m,c}|a=k) * P_{j-h_1}^t \end{cases}
\end{aligned} \tag{19}
$$

Where:

- $AF_{c,j-h_1}$ denotes the predicted allele fraction based on STR allele j at the locus for cell c.
- P_{j-h_1} denotes the phasing probability that STR allele j is linked with germline hSNP h_1.
- $P(r_{m,c}|a=j)$ denotes the read likelihoods given allele j based on the stutter error model.

M-step:

We calculated the phasing probabilities that quantify the phasing relationships between STR loci and hSNPs, which was initialized with 0.5. If STR allele j linked with germline SNP h_1 and allele k links with germline SNP h_2, the phasing probability given individual i read information D_i was calculated with the following equation:

$$
\begin{aligned}
&P^{t+1}(j-h_1, k-h_2|D_i) = \sum_{G_i^t} P^t(j-h_1, k-h_2|D_c, D_{bulk}, G_i^t) * P(G_i^t) \\
&= \sum_{G_i^t} \frac{P(j-h_1, k-h_2|G_i^t) * P(G_i^t) * P(D_{bulk}|j-h_1, k-h_2, G_i^t) * P(D_c|j-h_1, k-h_2, G_i^t)}{P(D_c, D_{bulk}|G_i^t)} \\
&= \sum_{G_i^t} \frac{P(j-h_1, k-h_2|G_i^t) * P(G_i^t) * P(D_{bulk}|j-h_1, k-h_2, G_i^t) * \prod_{c=1}^{n}[P(D_c|j-h_1, k-h_2, G_c^t=(j,k)) * P(G_c^t=(j,}{P(D_{bulk}|G_i^t) * \prod_{c=1}^{n}[P(G_c^t=(j,k)|G_i^t) * P(D_c|G_c^t=(j,k))]}
\end{aligned} \tag{20}
$$

Where:

- $P(G_i^t)$ denotes the population haplotype frequency prior.
- $p(D \mid j-h_1, k-h_2)$ is the genotype likelihood for individual i given allele j linked with SNP h_1 and allele k links with SNP h_2, and with $P(j-h_1, k-h_2|G_i^t)$ set to 0.5.
- $P(G_c^t = (j,k)|G_i^t)$ denotes the mosaic mutation probability that the cell genotype is (j,k) given G_i.

Similarly, the probability that STR allele j is linked with germline SNP h_2 follows the same approach. Of note, we only used reads covering both the candidate locus and the nearby hSNP to calculate the phasing probability.

The mosaic allele fraction for mutant allele $\neg k$ was calculated using the following function:

$$
\mu_{i,k\to\neg k}^{t+1} = \frac{\sum_{j\in A_i\setminus\{k\}} P(G_i^t=(j,k)) \sum_c P(G_i^t=(j,\neg k)|G_i^t=(j,k)) + P(G_i^t=(k,k)) \sum_c P(G_i^t=(k,\neg k)|G_i^t=(k,k))}{\sum_{j\in A_i\setminus\{k\}} P(G_i^t=(j,k)) \sum_c 1 + P(G_i^t=(k,k)) \sum_c 2} \tag{21}
$$

Where:

- $(G_i^t = (j,k))$ denotes the individual germline probability in E-step.
- k denotes the germline allele.

The sequence-based STR allele population frequency was calculated as follows:

$$f_j^{t+1} = \frac{1}{2N_{ind}} \sum_{i=1}^{N_{ind}} \sum_{k=1}^{A_i} \left(P\left(G_i^t = (j,k) \middle| G_{c,i}^t, D_{c,i}, D_{bulk,i}, \theta_{stutter}^t\right) + P\left(G_i^t = (k,j) \middle| G_{c,i}^t, D_{c,i}, D_{bulk,i}, \theta_{stutter}^t\right)\right) \quad (22)$$

Where:

- All parameters are defined in Equation (18).

Estimating somatic genotyping posteriors for each site across cells:

BayesMonSTR employed a dynamic programming approach to compute the posterior probability that a candidate locus contains at least one mutant cell. The posterior probability that a mosaic mutation occurs ($k > \neg k$) is calculated with the following formula:

$$P_{i,k\to -k} = 1 - P\left(L_{i,k\to -k} = 0 | \bar{D}_c\right) \quad (23)$$

Where:

- L_i denotes the mutant allele count for individual i.

The calculation follows with Bayes' rule:

$$\begin{aligned} P\left(L|\bar{D}_c\right) &= \frac{P(L) * P\left(\bar{D}_c|L\right)}{P\left(\bar{D}_c\right)} \\ &= \frac{P(L) * \sum_{\vec{G}_c} P\left(\bar{D}_c|\vec{G}_c, L\right) P\left(\vec{G}_c|L\right)}{\sum_{L'=0}^{N} P(L') \sum_{\vec{G}'_c} P\left(\bar{D}_c|\vec{G}'_c, L'\right) P\left(\vec{G}_c|L'\right)} \\ &= \frac{P(L) * \sum_{\vec{G}_c} P\left(\bar{D}_c|\vec{G}_c\right) P\left(\vec{G}_c|L\right)}{\sum_{L'=0}^{N} P(L') \sum_{\vec{G}'_c} P\left(\bar{D}_c|\vec{G}_c\right) P\left(\vec{G}_c|L'\right)} \\ &= \frac{P(L) * \sum_{\vec{G}_c} P\left(\bar{D}_c|\vec{G}_c\right) \delta_{L,S(\vec{G}_c)} \frac{1}{\binom{N}{L}}}{\sum_{L'=0}^{N} P(L') \sum_{\vec{G}'_c} P\left(\bar{D}_c|\vec{G}_c\right) * \delta_{L',S(\vec{G}'_c)}} \\ &= \frac{\frac{1}{\binom{N}{L}} * P(L) * \sum_{\vec{G}_c} \delta_{L,S(\vec{G}_c)} \prod_c P(D_c|G_c)}{\frac{1}{\binom{N}{L}} * \sum_{L'=0}^{N} P(L') \sum_{\vec{G}'_c} \delta_{L',S(\vec{G}'_c)} \prod_c P(D_c|G'_c)} \end{aligned} \quad (24)$$

The Kronecker delta function, denoted as $\delta_{L,S}$ is a mathematical function that evaluates to 1 when its indices are equal, and evaluates to 0 when its indices differ.

$$P(L_{i,k\to\neg k}) = \binom{N_c}{L_{i,k\to\neg k}} \mu_{i,k\to\neg k}^{L_{i,k\to\neg k}} (1 - \mu_{i,k\to\neg k})^{N_c - L_{i,k\to\neg k}} \quad (25)$$

The recursive formula of dynamic programming is as follows:

$$M_{l,j} = \begin{cases} \sum_{\vec{G}_c} P\left(\bar{D}_c|\vec{G}_c\right) \delta_{l,S(\vec{G}_c)} \prod_{c=1}^{j} P(D_c|G_c) & 0 \le l \le j \\ 0 & l > j \end{cases} \quad (26)$$

$$M_{l,j} = M_{l,j-1} * P(D_j|G_j = (j,k)) + M_{l-1,j-1} * P(D_j|G_j = (j,-k)) \quad (27)$$

Where:

- $M_{l,j}$ denotes probability for mutant allele count l and allele j in the DP matrix.

Haplotype phasing related artifacts

Haplotype phasing is widely adopted as a methodology to increase confidence of mosaic mutation calling[6]. The probabilistic haplotype phasing approach we designed is based on two key assumptions: 1) Given a relatively low mutation rate, it is highly unlikely that somatic mutations would occur independently on both haplotypes at the same position; 2) non-cancer cells possess diploid genomes.

Assuming STR allele j is linked with hSNP h_1, and STR allele k is linked with hSNP h_2. Discordant rate of reads for bulk sample d_{bulk} and single cell sequencing data d_{sc} were defined as

$$d_{bulk} = \frac{n_{k-h_1}}{n_{j-h_1} + n_{k-h_1}} \tag{28}$$

$$d_{sc} = \frac{n_{k-h_1}}{n_{j-h_1} + n_{k-h_1}} \tag{29}$$

Amplification error d_{amp} rate and discordant rate d_k for mutant cells sample were defined as:

$$d_{amp} = \frac{n_{j-h_2}}{n_{j-h_2} + n_{k-h_2}} \tag{30}$$

$$d_k = \frac{n_{k-h_1}}{n_{k-h_1} + n_{k-h_2}} \tag{31}$$

Where:

- n_{j-h_1} is the number of reads that STR allele j is linked with hSNP h_1.
- n_{k-h_1} is the number of reads that STR allele k is linked with hSNP h_1.
- n_{j-h_2} is the number of reads that STR allele j is linked with hSNP h_2.
- n_{k-h_2} is the number of reads that STR allele k is linked with hSNP h_2.

Based on empirical assessment, candidate mutations are considered high-confidence somatic events when all four of their phasing error rates (defined in Equations 28–31) are below 0.1.

Genome-wide variant prediction using Random Forest (RF) classification model

We trained two Random Forest classifiers to predict mosaic variants using a rigorously curated training set with phasing information (Supplementry Table 2). The first model incorporated phasing-related features, while the second model was designed for genome-wide prediction, regardless of phasability. A leave-one-donor-out strategy was applied in model training and testing.

The model was specifically trained to predict refined genotypes, categorizing sites into five distinct classifications: amplification error (amp error), sequencing error (seq error), repeat, heterozygous (het) and mosaic. To ensure an unbiased performance assessment, we employed a leave-one-out cross-validation strategy for model training and testing. To identify mosaic loci, we applied a dual-mode filtering strategy based on phasing feasibility. Phase-informed mode: Variants initially predicted by the phase Random Forest model underwent a series of phasing-based hard filters to yield a high-confidence set of phasing-corrected mosaic loci. These filters included thresholds for bulk, single-cell, and mutant-cell discordance rates, alongside an amplification error rate (all < 0.1, as defined in Equations 28–31), supplemented by requirements for mutant cell VAF ≥ 0.1 and a mutant read count ≥ 2. Non-phaseable mode: For variants lacking reliable

phase information, two sequential Random Forest models were applied to mitigate amplification errors. To derive a stringent set of true mosaic sites, we then applied an additional empirical filter requiring mutant cell VAF > 0.25 and mutant read count ≥ 3.

Examination of artifact sites identified by different software tools

To investigate the superior performance of BayesMonSTR over HipSTR and ExpansionHunter in detecting mosaic STR mutations from single-cell data, we systematically analyzed the false-positive calls from the latter two tools. This analysis revealed how BayesMonSTR's multi-step filtering pipeline eliminates such artifacts. The key filtering steps are defined as follows: 1. Prescan: Candidate loci lacking alternative alleles or residing in low-depth regions (mean depth ≤ 1 across cells) were removed. 2. Genotyping: Candidate mutations with a mosaic genotyping posterior probability below 0.9 were excluded. 3. Phasing: Candidate mutations were filtered based on the phasing metrics defined in Equations 28–31. Specifically, we excluded any mutation with a bulk discordant rate, single-cell discordant rate, amplification error rate, or mutant-cell discordant rate exceeding 0.1. 4. RF model: Candidate mutations were classified as mosaic if the highest predicted probability was for the mosaic class. The SHAP[7] (SHapley Additive exPlanations, SHAP, v0.46.0) method was then used to identify the key features driving the RF classification for these mutations (Extended Data Fig. 4e).

Simulation of spike-in mosaic STR mutations in single cell sequencing data

To simulate mosaic STR mutations in single cell sequencing data, STR indels and mismatches were randomly introduced at target loci using BAMSurgeon[8] (v1.4.1). Diploid single-cell data were then sampled via beta-binomial sampling to emulate allele-specific amplification imbalance. Finally, aligned reads from 10 single-cell libraries (male donors) were down-sampled and merged to generate synthetic diploid datasets for validation.

Breast cancer sample collection

A fresh breast cancer tissue sample and a matched normal breast tissue were obtained from a female breast cancer patient at The Second Affiliated Hospital Zhejiang University School of Medicine. The study was approved by the Institutional Review Board of Westlake University (IRB: 20230213DYM001). Informed consent was obtained from the patient in accordance with the IRB-approved protocol, and consent for data publication was also provided. At the time of collection, pathologist analysed the samples and categorized them into tumor and matched normal breast tissue blocks. Fresh tissue samples were transferred to sterile PBS buffer (Sangon Biotech, E607008) on ice and processed immediately upon arrival at the laboratory. The breast cancer and normal tissue samples were then dissociated, and viable single cells were isolated. Single-cell whole genome amplification, library preparation, and whole-genome sequencing were subsequently performed as described below.

Isolation of single cells of breast tissue samples for scWGS

Fresh tumor and matched normal breast tissue samples were collected and processed within two hours post-surgery. For each sample, small tissue blocks of approximately

1096 1 cm^3 were isolated for dissociation. The tissue dissociation procedure was performed
1097 according to previously described methods[9,10], with slight modifications. Briefly, tissue
1098 samples were manually minced in 1.5 mL tubes containing collagenase type 3 (300
1099 U/mL) (Worthington Biochemical, #LS004182) and hyaluronidase (100 U/mL)
1100 (Worthington Biochemical, #LS002592) for 2 hours, with gentle pipetting every 60
1101 minutes. The digested tissue was then centrifuged at 1500 rpm for 5 minutes. The pellet
1102 was digested with 5 mL of pre-warmed 0.25% trypsin-EDTA (Thermo #25200072) at
1103 37°C for 5 minutes. Digested cells were washed once with HBSS + 2% FBS (Jackson,
1104 #017-000-121), then treated with 0.1 mg/mL DNase I (Worthington Biochemical,
1105 #LS002139) in DMEM/F12 media at 37°C for 5 minutes. After washing, the cells were
1106 filtered through a 40 μm mesh. Dissociated cells were counted and resuspended in the
1107 designated medium for subsequent staining and FACS sorting. Meanwhile, tissue
1108 surrounding the small blocks (within 0.5 cm from the edge) was similarly dissociated
1109 into a single-cell suspension. DNA was then extracted using a DNeasy Blood & Tissue
1110 Kit (Qiagen, 69504) from these suspensions for bulk DNA whole genome sequencing.
1111 For FACS sorting, single cells were stained with the following antibodies: CD45
1112 (BioLegend, 304022), CD31 (BioLegend, 303114), CD49f (Thermo, 12-0495-82),
1113 EpCAM (BioLegend, 324204), and CD34 (STEMCELL Technologies, 60013Az.1).
1114 The antibodies were used at a dilution of 1:200, with appropriate conjugated
1115 fluorophores, and incubated for 30 minutes on ice. After staining, cells were washed
1116 and resuspended in Hanks Balanced Salt Solution (HBSS) + 2% FBS + Antibiotic-
1117 Antimycotic + 4′,6-diamidino-2-phenylindole (DAPI) (1 μg/mL) at a density of 5
1118 million cells/mL. The stained samples were then analyzed and sorted on a FACSAria
1119 Fusion (BD Biosciences) with a 100-μm nozzle. The target single cells were sorted into
1120 individual tubes, each containing a single cell and 3 μL of cell buffer, prepared for
1121 subsequent single-cell whole genome amplification. The FACS data were further
1122 processed using FlowJo V10. Gating strategy is shown in Extended Data Fig. 4.

1123 **Whole genome amplification and single cell DNA & RNA sequencing**

1124 The sorted single cells were subjected to whole-genome amplification using either
1125 REPLI-G (Qiagen, 150023) or ResolveOME (Bioskyrb). Single cells amplified with
1126 ResolveOME yielded both whole-genome amplification products and full-length
1127 transcriptome amplification products. In some cases, both reactions were combined to
1128 generate both whole-genome amplification products (via REPLI-G) and full-length
1129 transcriptome amplification products from the same single cell. All whole-genome and
1130 full-length transcriptome amplification products from ResolveOME underwent library
1131 preparation using the ResolveOME library preparation protocol, including a 9-10 cycle
1132 PCR library amplification. For REPLI-G whole-genome amplification products, as well
1133 as the extracted paired bulk DNA, PCR-free library preparation was performed using
1134 the Hieff NGS Ultima Pro PCR Free DNA Library Prep Kit V2 (Yeashen, 12196).
1135 The WGS libraries were sequenced on the BGISEQ DNBSEQ-T7 platform (BGI Inc.)
1136 using a paired-end sequencing length of 150 bp (PE150). The average sequencing depth
1137 for each WGS library was 30–40x, while the bulk WGS library had an average

1138 sequencing depth of 200x. Full-length transcriptome libraries were sequenced on the
1139 NovaSeq X Plus platform (Illumina Inc.) using the PE150 strategy, generating
1140 approximately 20 million reads per library. Additionally, parallel library preparation of
1141 the same whole-genome amplification product or parallel sequencing of the same
1142 library was conducted to obtain WGS data duplication.

1143 **Processing of RNA sequencing data**

1144 For RNA sequencing data, trimmomatic[11] (v 0.39) was first used to remove adapters.
1145 Salmon[12] (v 1.10.2) was used to quantify the expression of transcripts. The salmon
1146 index was constructed using the GTF and FASTA files (v GRCh37.87) from
1147 ENSEMBL.

1148 **Comparison of algorithms detecting mosaic STR mutations from the**
1149 **benchmarking dataset**

1150 To adapt the two germline mutation callers for identifying mosaic STRs from single
1151 cell sequencing data, we applied a series of stringent filters to each software to generate
1152 a mosaic mutation call set.

1153 HipSTR[3] (v0.6.2) was run using the mode of "*de novo* stutter estimation + STR calling
1154 with *de novo* allele generation", variant calls were then filtered following recommended
1155 empirical cutoffs (--min-call-qual 0.9 --max-call-flank-indel 0.15 --max-call-stutter
1156 0.15 --min-call-allele-bias -2 --min-call-strand-bias -2). ExpansionHunter[13] (v5.0.0)
1157 with default settings were used for variant callings of benchmark datasets.

1158 Germline genotypes and mosaic genotypes were identified as the most frequent and
1159 second most frequent genotypes across all cells. Cell-specific variants were identified
1160 by those that have one mutant cell, and shared variants were identified as those with
1161 multiple mutant cells. In addition, if there are less than five valid cell calls, the candidate
1162 loci were excluded from the mutation list.

1163 For BayesMonSTR, variants were identified using a genotyping cutoff of mosaic
1164 posterior $\geqslant$ 0.9, predicted as mosaic by the RF model, and passing a series of hard
1165 filters. More specifically, for phasable sites, the hard filters include a series of phasing-
1166 related filters (equations 28-31, bulk discordant rate < 0.1, single cell discordant rate <
1167 0.1, amplification error rate < 0.1 and mutant cell discordant rate < 0.1), and other filters
1168 (VAF $\geqslant$ 0.1 and mutant read count $\geqslant$2); For non-phasable sites, the hard filters
1169 include VAF > 0.25 and mutant read count $\geqslant$ 3.

1170 **Cross-validation of candidate mutations**

1171 To evaluate candidate loci identified by different software tools, we assessed their
1172 performance using orthogonal data: duplicate libraries from the same single cell, RNA
1173 sequencing data from the same cell, and bulk sequencing data from adjacent tissues.

1174 In both duplicate library sequencing data and single-cell DNA&RNA data, a one-tailed
1175 binomial test (lower-tail) was applied, where n represents the total read count and k
1176 represents the mutant read count. For RNA-based validation, $\theta = 0.5$ was used as the
1177 null hypothesis parameter. For duplicate cell validation, the test was conducted using
1178 the maximum p-value derived from tests with either the VAF in the mutant cell or $\theta =$
1179 0.5 as the null hypothesis parameter. Validation outcomes were classified as validated

1180 ($p \geq 0.05$ and $k > 0$), unvalidated ($p < 0.05$ and $k = 0$), or undetermined (all other cases).
1181 Validation rates were calculated after excluding undetermined loci.
1182 Shared variants were further evaluated by manual inspection and PhyloSOLID (v 0.0.1)
1183 STR genome validation mode (https://github.com/douymLab/PhyloSOLID). Cell
1184 genotype likelihoods along with alternative and total read depth information were used
1185 for shared variants evaluation based on phylogenetic tree.
1186 We also assessed the mutations from DNA utilizing RNA sequencing data from the
1187 same cell (Extended Data Fig. 4a). This orthogonal validation approach enabled us to
1188 evaluate the concordance between DNA-based mosaic detection and transcript-level
1189 variation.

1190 **Visualization of STR mutations**

1191 INSIGHT (v 0.0.1) developed by our lab were used for visualization of STR mutations
1192 (https://github.com/douymLab/INSIGHT).

1193 **Identification of mosaic SNVs from single cells**

1194 Similar strategies were applied to detect mosaic SNVs in the same datasets. After
1195 sequence alignment and data preprocessing, raw candidate loci were identified by
1196 mpileup in SAMtools[14] (v1.13) with the parameters '-B -d 8000 -q 10 -Q 10'. Then, the
1197 mutations were identified by using MosaiSC-DNA[15] (v0.0.1,
1198 https://github.com/douymLab/mosaicSC). The predicted SNVs were further filtered
1199 with mutant cell VAF $\geqslant$ 0.3 and mutant read count $\geqslant$ 3 to get a more reliable set of
1200 mosaic SNVs.

1201 **Phylogeny reconstruction using shared variants**

1202 Given the modest validation rate for shared variants, the constructed phylogenetic tree
1203 (Figure 2g) was based on manually checked cell-by-mutation matrix (Supplementry
1204 Table 5). PhyloSOLIDD[16] (v0.0.1) was used for phylogeny reconstruction.

1205 **Correlation of age with multiple mutation-derived features**

1206 We performed association analyses between age and multiple mutation-derived
1207 features, including mutation rate, variant length, and cell proportion with >5bp indels.
1208 Among these, the mutation rate was calculated by dividing the number of mutations by
1209 the total length of callable regions. We used scipy[17] (v 1.14.1) to perform linear
1210 regression analysis between age and features.

1211 **Mutation signature analysis**

1212 For mutations in STR regions, mutation matrices were generated using
1213 SigProfilerMatrixGenerator[18] (v 1.2.31), and then, SigProfilerAssignment[19] (v 0.1.9)
1214 was used to assign known COSMIC mutation signatures to every individual, which was
1215 accomplished using the forward stagewise algorithm and nonnegative least squares. In
1216 our samples, three 83-cat signatures of small insertions and deletions (ID1, ID2, ID12)
1217 and multiple 96-cat signatures of single base substitutions were found to be active.
1218 Additionally, SigProfilerAssignment was used to assign signature probabilities to every
1219 mutation, allowing to quantify the mutation burden associated with each signature for
1220 every individual.

Additionally, we also used a redefined InDel taxonomy[20] to illustrate mutation profile for InDels in STR regions, and the mutation signature similarity was computed via cosine similarity was used to compute similarity between our InDels and those caused by the *MSH2* and *MSH3* knockouts[20].

Specially, for SNVs in STR regions, a custom method was developed to classify SNVs into different categories. Initially, mutations were distinguished based on the six kinds of fundamental base changes, followed by classification according to the nucleotide context of the entire STR region in which the SNV occurred (considering both repeat unit length and total length of the repeat region). Subsequently, non-negative matrix factorization was performed using MutationalPatterns[21] (v 3.12.0) to extract three distinct *de novo* mutation signatures. Based on the relative contribution of each signature, the mutation burden associated with each signature for every individual was quantified.

Relationship between mosaic SNVs in DNA repair gene and mosaic STR indel burden

The list of DNA repair genes was obtained from the comprehensive compilation maintained by the laboratory of Richard D. Wood at The University of Texas MD Anderson Cancer Center (https://www.mdanderson.org/research/departments-labs-institutes/labs/wood-laboratory/resources.html). This curated gene set represents an updated synthesis based on previously published reviews.

ANNOVAR[22] (v 2020-06-07) was used to identify exonic variants, their associated genes, and their effects on the protein sequence, including synonymous, nonsynonymous, stop-gain, stop-loss, and frameshift variants. Variants annotated as "unknown" by ANNOVAR due to incomplete ORF information were re-annotated using the Ensembl Variant Effect Predictor[23] (VEP, https://www.ensembl.org/vep) to determine their coding consequences based on updated transcript models.

To assess the association, we fitted a generalized linear mixed model (GLMM) with a negative binomial distribution to account for overdispersion in count data. The model included somatic SNV status (presence in DNA repair gene vs. absence) as a fixed effect and a random intercept for individual to account for within-individual correlations (e.g., multiple neurons or samples per individual). The significance of fixed effects was evaluated using Wald tests. All analyses were performed in R (v 4.3.3) using the glmmTMB[24] (v 1.1.14) package.

Identification of potentially functional-disrupting mosaic STR mutations

ANNOVAR was employed to determine the nearest gene of mosaic STR mutations and identify potentially damaging exonic variants, including nonsynonymous, stopgain, stoploss, and frameshift variants. CADD[25] (v 1.7) scores were incorporated to assess variant deleteriousness. ENCODE's candidate cis-regulatory elements annotation[26] was used to identify variants with potential regulatory functions. The chromatin state annotation from ENCODE was used to identify variants in active TSS and enhancer regions. Chromatin state annotations were derived from the ChromHMM 18-state

model from the ENCODE portal[27] for seven primary B cell samples, 17 lung tissues and 16 brain tissues.

Potentially functional-disrupting STR indels were defined as variants in upstream, UTR or exonic regions annotated by ANNOVAR, or regions defined as chromatin state of active TSS and enhancer by ENCODE, or variants have CADD score greater than 20. Gene affected by a STR indel was determined as the nearest gene annotated by ANNOVAR.

Identification of mosaic STR mutations in functional non-coding elements

To identify STR mutations may have a significant impact on gene expression, we utilized the human G-quadruplexes prediction panel from EndoQuad[28] for G-quadruplex forming regions, the Level 2 R-loop annotation panel from R-loop Base[29] for R-loop forming regions, ENCODE's candidate cis-regulatory elements annotation[26] for CTCF-binding sites, and the CpG islands annotation from the UCSC Genome Browser[30] for CpG islands.

Identification of mosaic STR mutations near genes of high expression

Gene expression data of whole blood, lung and brain cortex tissues was downloaded from GTEx[31]. Within each tissue, we calculated quartiles of expression of all genes near STR loci and categorized each gene into four categories (Q1 to Q4) based on average expression level across samples, with Q1 representing expression level is in minimum to first quartile.

Gene ontology (GO) and disease enrichment

KOBAS[32] (v 3.0) was used to conduct Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) disease enrichment of genes near STR indels. To be specific, Fisher's exact test was used to calculate P-values, while Benjamini and Hochberg (1995) method was used to control the FDR.

Process of snATAC-seq data

Raw FASTQ files of snATAC-seq were aligned to the hg38 reference genome, using 10x cellranger-atac count toolkit (v 2.1.0). Only cells identified through cellranger's cell calling were included in subsequent analyses. Duplicate reads were first filtered to ensure data integrity. Reads sharing identical cell barcodes and genomic coordinates (start and end positions) were grouped. Within each group, only reads with identical sequences were considered duplicates; among these, the read with the highest mapping quality and average base quality was retained.

Detection of mosaic STR mutations from snATAC-seq

Building upon the framework of BayesMonSTR, we adapted the approach to better handle the sparsity of single-cell ATAC-seq data by performing genotyping at the pseudo-bulk level.

The stutter error model was trained using 26 snATAC-seq datasets derived from brain tissue and whole blood samples, obtained from the publicly available studies SRP309633 and SRP192525. After estimating the stutter error model and extracting alleles through STR segmentation as mentioned before, we directly estimated individual-level mosaic posterior probabilities at the pseudo-bulk level.

We than perform multiple strict filters to retain highly confident mutations including

1. posterior probability exceeding 0.5.
2. count of used reads in genotyping exceeding 10.
3. homozygous in the bulk genotype with exactly one alternative read.
4. no other reads at the locus contained off-target stutter artifacts or mismatches.
5. recurred in less than two individuals.

Per-cell mutation rate estimation

In snATAC-seq data, the mutation rate per cell can be estimated for each individual cell as the ratio of the number of detected mutations to the size of its callable region. Callable regions were defined as regions covered by at least one read spanning STR with high mapping quality. And the mutation rate of an individual (μ) can be estimated by the mean of cells' mutation rate:

$$\mu = \frac{\sum_{i=1}^{N_c}(\frac{M_i}{L_i})}{N_c} \tag{32}$$

Where:

- M_i: the number of somatic mutations detected in cell i;
- L_i: the total length (bp) of the callable genomic regions (i.e., sequenced with sufficient coverage and quality) in cell i;
- N_c: the total number of cells analyzed.

The mutation rate of the regulatory region was calculated using the same strategy, with the exception that regulatory callable regions were used instead of whole callable regions.

Dual-Luciferase promoter reporter assay

The promoter sequences of *NRN1* and *CNTNAP2* were PCR-amplified from HEK293T cDNA templates and cloned into our dual-luciferase reporter plasmid vector. All constructs were verified by sequencing. 0.2M HEK293T cells were seeded into 24-plate well with DMEM for 12 hours, cells were transiently transfected with plasmids using polyethyleneimine transfection reagent (1mg/mL; Polyscience, #23966-1). After 48 hours, promoter activity was analyzed using the Dual-Luciferase Reporter Gene Assay Kit (Yeasen Biotechnology, 11402ES60) following manufacturer's protocol on a Varioskan LUX Microplate reader (Thermo). Relative promoter activity was calculated as the ratio of firefly/Renilla luminescence.

Data availability

WGS data and single-cell RNA-seq data from breast cancer samples generated in this study will be publicly available in the Sequence Read Archive (SRA). WGS data for neurons, B cells, and lung epithelium are accessible through dbGaP under accession numbers phs001485.v3.p1 and phs001808.v1.p1, and in the European Genome-phenome Archive (EGA) under accession EGAD00001005193, respectively. sn-ATAC-seq data from brain tissue are available in the EGA (accession EGAS00001001934) and Synapse (accession syn52293417).

Code availability

BayesMonSTR is implemented in Python and R and is licensed under the MIT License. The source code, documentation and examples are available on GitHub at https://github.com/douymLab/BayesMonSTR.

References

1. Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* 1–11 (2023) doi:10.1038/s41586-023-06045-0.

2. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).

3. Willems, T. *et al.* Genome-wide profiling of heritable and de novo STR variations. *Nat Methods* **14**, 590–592 (2017).

4. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).

5. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

6. Dou, Y. *et al.* Accurate detection of mosaic variants in sequencing data without matched controls. *Nat Biotechnol* **38**, 314–319 (2020).

7. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).

8. Ewing, A. D. *et al.* Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat Methods* **12**, 623–630 (2015).

9. Yuan, L. *et al.* Reconstruction of dynamic mammary mini gland in vitro for normal physiology and oncogenesis. *Nat. Methods* **20**, 2021–2033 (2023).

10. Fu, A. *et al.* Tumor-resident intracellular microbiota promotes metastatic colonization in breast cancer. *Cell* **185**, 1356-1372.e26 (2022).

11. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

12. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

13. Dolzhenko, E. *et al.* ExpansionHunter: A sequence-graph based tool to analyze variation in short tandem repeat regions. *Bioinformatics* **35**, 4754–4756 (2019).

14. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

15. Zheng, Y., Fan, W., Lu, J., Yang, Q. & Dou, Y. Accurate detection of specific and shared mosaic variants in single-cell whole-genome sequencing data in normal human cells. (2025).

16. Yang, Q. *et al.* PhyloSOLID: robust phylogeny reconstruction from single-cell data despite pervasive errors and extreme sparsity.

17. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

18. Bergstrom, E. N. *et al.* SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genom.* **20**, 685 (2019).

19. Díaz-Gay, M. *et al.* Assigning mutational signatures to individual samples and individual somatic mutations with SigProfilerAssignment. *Bioinformatics* **39**, btad756 (2023).

20. Koh, G. C. C. *et al.* A redefined InDel taxonomy provides insights into mutational signatures. *Nat. Genet.* **57**, 1132–1141 (2025).

21. Manders, F. *et al.* MutationalPatterns: the one stop shop for the analysis of mutational processes. *BMC Genom.* **23**, 134 (2022).

22. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164–e164 (2010).

23. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

24. Brooks, M., E. *et al.* glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *R J.* **9**, 378 (2017).

25. Schubach, M., Maass, T., Nazaretyan, L., Röner, S. & Kircher, M. CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. *Nucleic Acids Res.* **52**, D1143–D1154 (2024).

26. Abascal, F. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

27. Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.* **48**, D882–D889 (2019).

28. Qian, S. H. *et al.* EndoQuad: a comprehensive genome-wide experimentally validated endogenous G-quadruplex database. *Nucleic Acids Res.* **52**, D72–D80 (2023).

29. Lin, R. *et al.* R-loopBase: a knowledgebase for genome-wide R-loop formation and regulation. *Nucleic Acids Res.* **50**, D303–D315 (2021).

30. Perez, G. *et al.* The UCSC Genome Browser database: 2025 update. *Nucleic Acids Res.* **53**, D1243–D1249 (2024).

31. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

32. Bu, D. *et al.* KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Res.* **49**, W317–W325 (2021).